(continued from part 25)

Other types of aerial

At very low frequencies, the length of a dipole aerial (otherwise known as a **half-wave** or **Hertz aerial**) will become impractically long. It was Marconi who found that the length of a dipole aerial can be halved by earthing one end of it. The **Marconi** (or **earthed**) **aerial** is therefore a vertical earthed mast, the height of which being determined as one quarter of the wavelength of the signal to be broadcast or received. The operating theory behind this aerial is that the ground itself acts as the bottom half of a dipole, providing a mirror image of a quarter wavelength. This aerial is, of course, vertically polarized.

There are many other types of aerial ranging from simple lengths of wire, through to complex arrays (similar to those used for domestic television reception) that utilise the concept of the **folded dipole** and **loop aerials**. There are also dish aerials which make use of a parabolic reflector to concentrate the received electromagnetic waves onto a single point. Each type of aerial is specifically designed to deal with the reception and transmission of electromagnetic radiation of various frequencies.

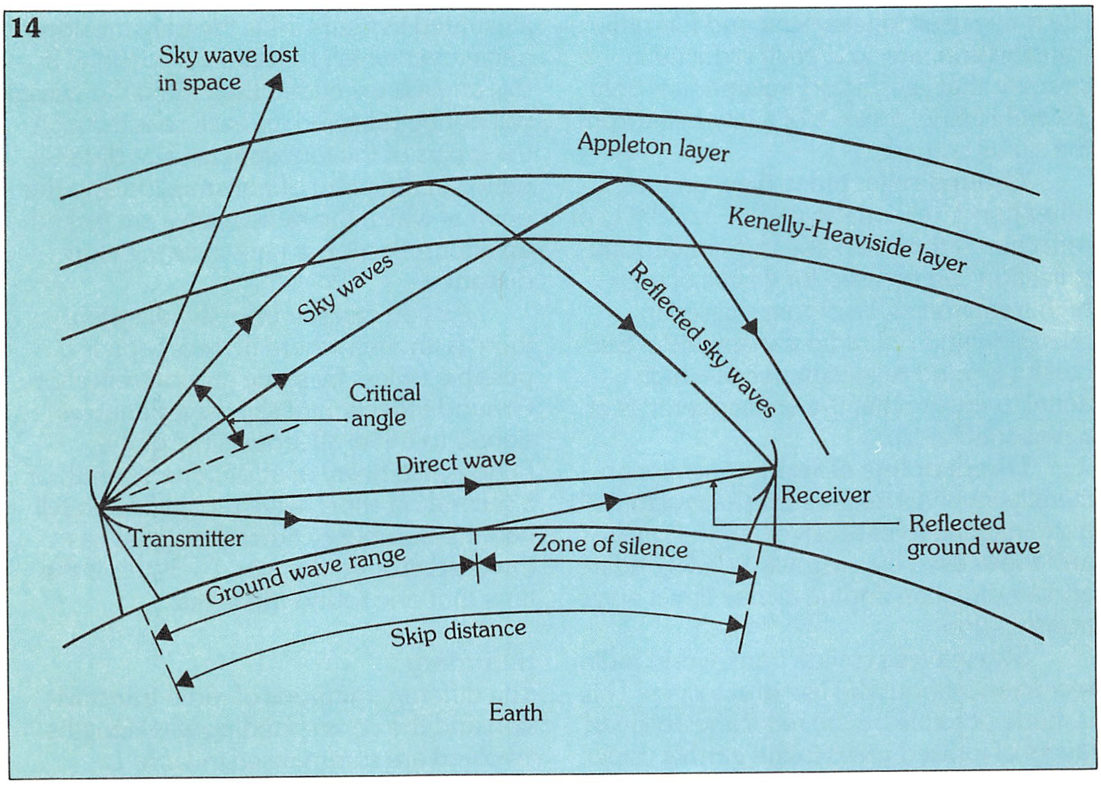
Electromagnetic wave propagation

Two factors affect the path and distance travelled by the radiated electromagnetic wave. The first is the weather, because it affects the atmosphere through which the waves travel, and the second is the carrier frequency of the radiated waves. Returning to the torch analogy, we know that it is difficult to obtain a well directed beam of light on a foggy night – even using a powerful beam – because the light energy is reflected by the particles of moisture in the air.

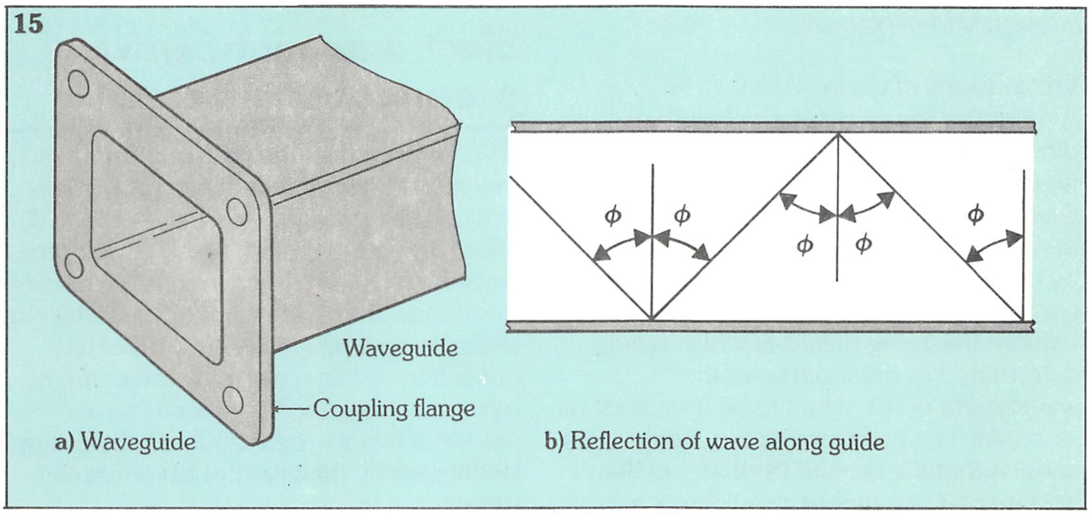
Radio waves have similar problems with rain, snow, sleet and other weather effects, being reflected, losing energy, or becoming dissipated, depending on the atmospheric conditions and the frequency of the signal.

Ground, direct and sky waves

Radio waves can travel from their transmitters to their reception points in a number of ways, the most important of which are known as: the ground or surface wave; the ground reflected wave; the direct or line of sight wave; and the sky wave. All of these



14. Transmission of electromagnetic waves through the atmosphere.



15. (a) Waveguide; (b) transmission of wave along the guide.

are illustrated in *figure 14*.

Ground or surface waves travel slightly above the earth's surface and are affected by the characteristics of the ground over which they travel. These waves are usually effective at low frequencies, as they are increasingly absorbed as the frequency rises. These so-called 'long waves' are therefore used in ship to shore communications and have a wavelength of more than 1000 m enabling them to be reliably received up to 1000 miles from their point of transmission.

Ground waves at high frequencies, like those used in television and FM radio transmission, are so greatly reduced in strength that satisfactory reception is only possible within an area of a few miles from the signal source.

Ground reflected waves, on the other hand, rely on the earth's capability of reflection rather than absorption. Ground reflection can be used for waves of very high frequencies, like those utilised in television and FM radio transmission. Successful ground reflection needs a high aerial to ensure that the area covered is of a reasonable size.

Direct or line of sight transmission literally involves transmitting a waveform in a straight line to its receiver. UHF signals and those used by microwave links, radar and air direction finding all use direct wave transmission.

Sky waves provide trans-world radio reception without the use of satellites. This is made possible by the reflecting ability of layers of ionised gases in the earth's upper atmosphere. Several of these layers have been discovered: the **Kennelly-Heaviside layer** is responsible for reflection in the daylight hours; and the **Appleton layer** permits signals to be reflected at night. A typical transmission using the Kennelly-Heaviside layer would be a distance of 1500 miles with a frequency greater than 20 MHz.

Sky wave communication is inconsistant but generally predictable, with the properties of reflection varying with sun spot activity and seasonal and yearly weather variations. The **critical angle** illustrated in *figure 14* is the transmission angle over which reflection is possible. If this angle is exceeded then the radio waves will not be reflected and will pass through the layers of the ionosphere. The distance between the point of transmission and the point at which the radio waves are picked up after reflection is known as the **skip distance**.

As you can see from the diagram, there is an area where no reception is possible unless by some ground reflection. Ground reflection of sky waves enables signals to travel all around the globe. Echoes at a delay of about one seventh of a second on short wave reception can tell you the number of times the signal has travelled around the world – as this is the time that one round trip takes.

Summary

The different methods of radio transmission and the corresponding wavelengths involved are summarised in *table 1*.

Waveguides

Instead of being transmitted into space, ultra-high frequency electromagnetic waves known as **microwaves** can also be propagated along hollow conductors known as **waveguides**: The transmitted wave is reflected along the inside of the waveguide as shown in *figure 15*. Waveguides provide private, stable transmission media that are usually used as feeders to and from microwave aerials, or in underground microwave telephone link systems.

Waveguides are restricted to the microwave band as the cross-sectional dimensions of the guide are related to the wavelength in much the same way that the dimensions of an organ pipe relate to the frequency of the note produced. A waveguide that can be used to carry up to 250,000 telephone channels will operate on a frequency between 40 and 100 GHz and will be around 5×10^{-2} m in size.

In such a system, the atmosphere inside the guide will be controlled to contain very few oxygen atoms, thus enabling the energy transmitted to travel down the pipe with little loss. This is a significant advantage because the range of transmission can be improved, with repeater stations being needed only every 30 km or so, instead of the 5 to 10 km previously required. A waveguide system operating at 40 GHz will need only a tenth of the repeater stations required by an open air broadcast system operating at the same frequency.

Another advantage of waveguides is that no outside or unwanted electromagnetic energy can interfere with the information being transmitted. All external radiation is blocked by the waveguide's metal walls, thus ensuring a reliable and dependable communications link.

The electromagnetic signal is propagated into and collected from the waveguide by various methods by using a system of small metal rods arranged to act as aerials. It is important to remember that the signal is transmitted mainly through the atmosphere in the centre of the guide, although part of the energy is conveyed by the associated currents and changes in the walls. This is, of course, because we are handling an electromagnetic wave and not an electric current.

Table 1
Frequency band subdivisions

Band number	Frequency range	Frequency/wavelength subdivision	Uses
2 3	30 – 300 Hz 300 – 3000 Hz	ELF – extremely low frequency VF – voice frequency	
4 5	3 – 30 kHz 30 – 300 kHz	VLF – very low frequency LF – low frequency	Initially used for radio telegraphy, now used for specialist applications such as ship to shore communication
6 7	300 – 3000 kHz 3000 – 30,000 kHz	MF – medium frequency HF – high frequency	Commercial AM broadcasting short wave and amateur operations. Good ionsphere reflection properties
8 9	30 – 300 MHz 300 – 3000 MHz	VHF – very high frequency UHF – ultra high frequency	Line of sight communications – televisions, FM radio and mobile communications – all on local level
10 11	3000 – 30,000 MHz 30,000 – 300,000 MHz	SHF – super high frequency EHF – extra high frequency	Radar radio relay, navigator, experimental work
12	300,000 – 3,000,000 MHz		Used for wideband communications, radar and microwave cookers

Coaxial cables

As we have seen, electromagnetic waves are propagated by a changing electrical current fed into an aerial. How, then, does the current not cause electromagnetic waves to be radiated from the cable that connects the aerial to the transmitter? The answer lies in **coaxial cable**.

As you can see from *figure 16*, this type of cable comprises two conductors – one inside the other – *both sharing the same concentric axis*. The dielectric that separates the two conductors can be either made of polythene or air – in the latter case, the two conductors are held apart with polythene disks. The outer conductor is earthed (set to 0 V), with the signal being connected to the centre conductor: the outer conductor therefore **screens** the inner, thereby reducing the effect of **capacitive pick-up.**

Capacitive pick-up has been covered in *Digital Electronics 22*, but just to recap, look at *figure 17*. Here, a wire carries a signal from a source to a receiver and an earth return is used to complete the circuit. Another wire, running close to the signal cable, carries a different signal. This unwanted signal, however, interferes with the source signal, and this interference is caused by capacitive pick-up.

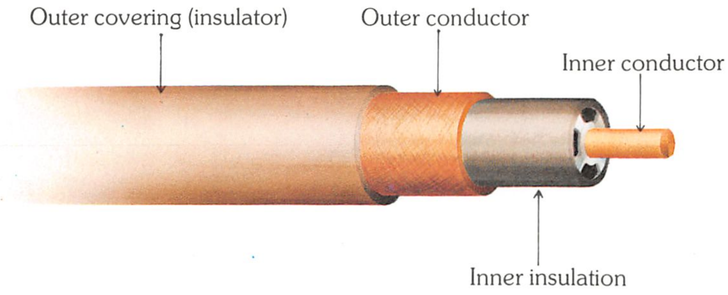
Looking at the diagram again, you can see that a capacitance, C_X, exists between the signal wire and the interfering wire, and that a capacitance, C_Y, exists between the signal wire and earth. (There is also a capacitance between the interfering wire and earth, however, this is of no consequence here.) These two capacitances act as a potential divider network, which feeds part of the interfering signal to the source signal wire.

However, using coaxial cable (*figure 18*), the adjacent wire is not able to affect the source signal as the capacitive effect only exists between the wire and the outer coaxial conductor – from here it drains to earth. This represents the major advantage of coaxial cable over ordinary cabling.

The multiplexed messages in telephone systems are transmitted with the aid of carrier frequencies. These 'carry' the speech signals 'piggyback fashion', in much the same way as the carrier frequencies used in A.M. radio transmission.

At these high frequencies (1 MHz or above) what is known as the **skin effect**

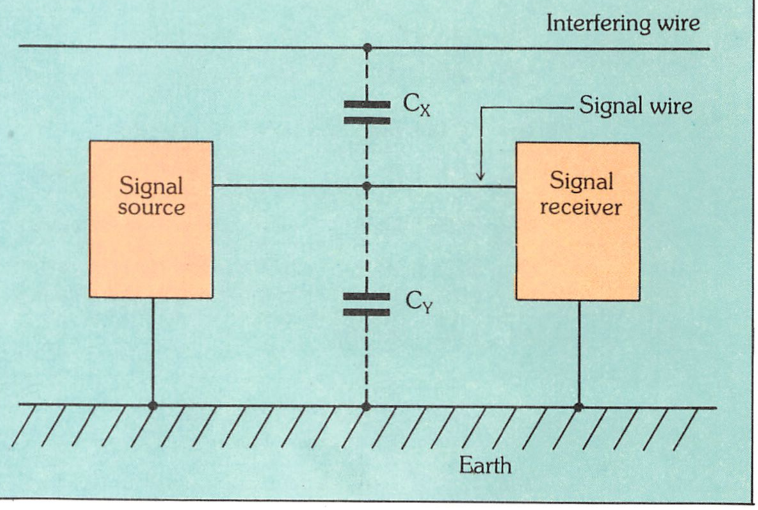
16



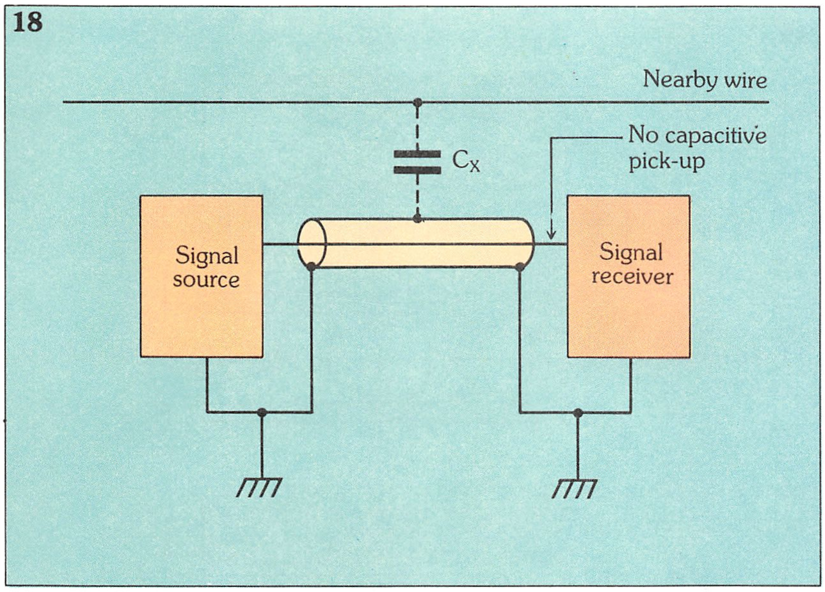
16. Coaxial cable.

17. Capacitive pick-up causes interference of the source signal by an unwanted signal from an adjacent wire.

17



18



18. With coaxial cable there is no capacitive pick-up.

becomes significant. At high frequencies, currents flowing in conductors are confined mainly to their outer surfaces. In a coaxial cable, the current flowing in the inner conductor flows on the outside of the wire, while the current in the outer conductor or screen is confined to the inside of that wire. Even if the outer conductor is not earthed, any capacitive pick-up from other cables is confined to the *outside* of the screen. This ensures that the effect of capacitive pick-up on the current flowing on the inside surface of the conductor is significantly less when high frequency signals are involved.

However, at lower frequencies, signal currents are not confined to the wire surface and flow right through the conductors: the isolating effects of the skin effect therefore do not occur. The use of screening, though, is still successful, as long as both ends of the outer conductor are held at 0 V – this is effective for reasonably short cable lengths, say a few hundred metres or so. For kilometre lengths of cable, however, the resistance of the outer conductor means that while the ends may be held at 0 V, the bulk of the conductor inbetween will not be earthed. Signals will then appear on the outer conductor, to be picked up by any adjacent cable's screen and then transferred to its inner conductor, resulting in cross-talk. Coaxial cable is therefore ideal for carrying *high frequency* information in communications systems, as the high frequencies reduce interference and cross-talk by utilising the skin effect.

Fibre optics

Optical transmission, one of the newest developments in communications technology, is made possible by very fine circular strands of glass – the optical fibres. Information, in the form of infra-red and visible light, can be propagated along these optical fibres by a phenomenon known as **fibre optics**. *Figure 19* illustrates cross-sections of three types of fibre optic strands.

Communications can be effected by transmitting information through optical fibres in the form of digitally coded pulses of light, usually provided by gallium arsenide light emitting diodes or injection lasers. Both of these are signal frequency light sources, the laser, as it is the more powerful device, being used for longer distance communication.

Signal reception is carried out by means of an avalanche or a p-i-n photodiode (see *Solid State Electronics 22*) connected to a suitable decoding and receiving circuit.

Optical fibres comprise two types of glass, arranged in a similar manner to the coaxial cable discussed earlier. The central core is composed of a strand of very pure low-loss glass. The outer covering consists of a similar type glass but has a slightly lower refractive index, ensuring that any light transmitted down the central core is subject to total internal reflection, and is not lost through the fibre's outside wall.

As you can see from *figure 19*, the three types of optical fibre possess different light propagation characteristics which can affect the accuracy of transmitted information. The internal design of the fibre is particularly important – any splitting of the light into its constituent wavelengths (as through a prism) has a detrimental effect on the shape of the pulse being transmitted and hence the nature of the information. Fibres which have a narrow core are usually chosen (*figure 19c*) as this arrangement presents a suitable medium for monochromatic light propagation – the type of light supplied by lasers and gallium arsenide LEDs.

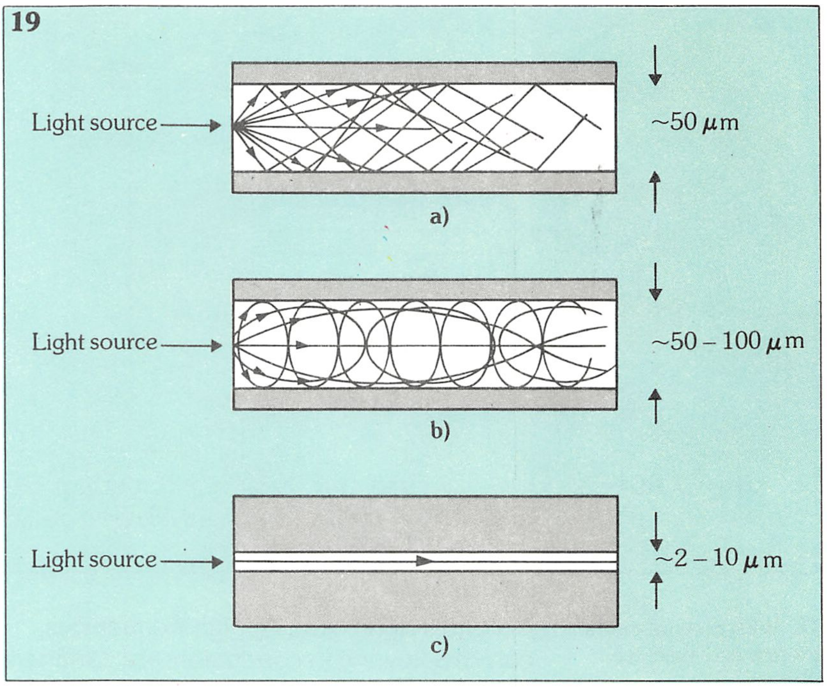
Advantages of fibre optics

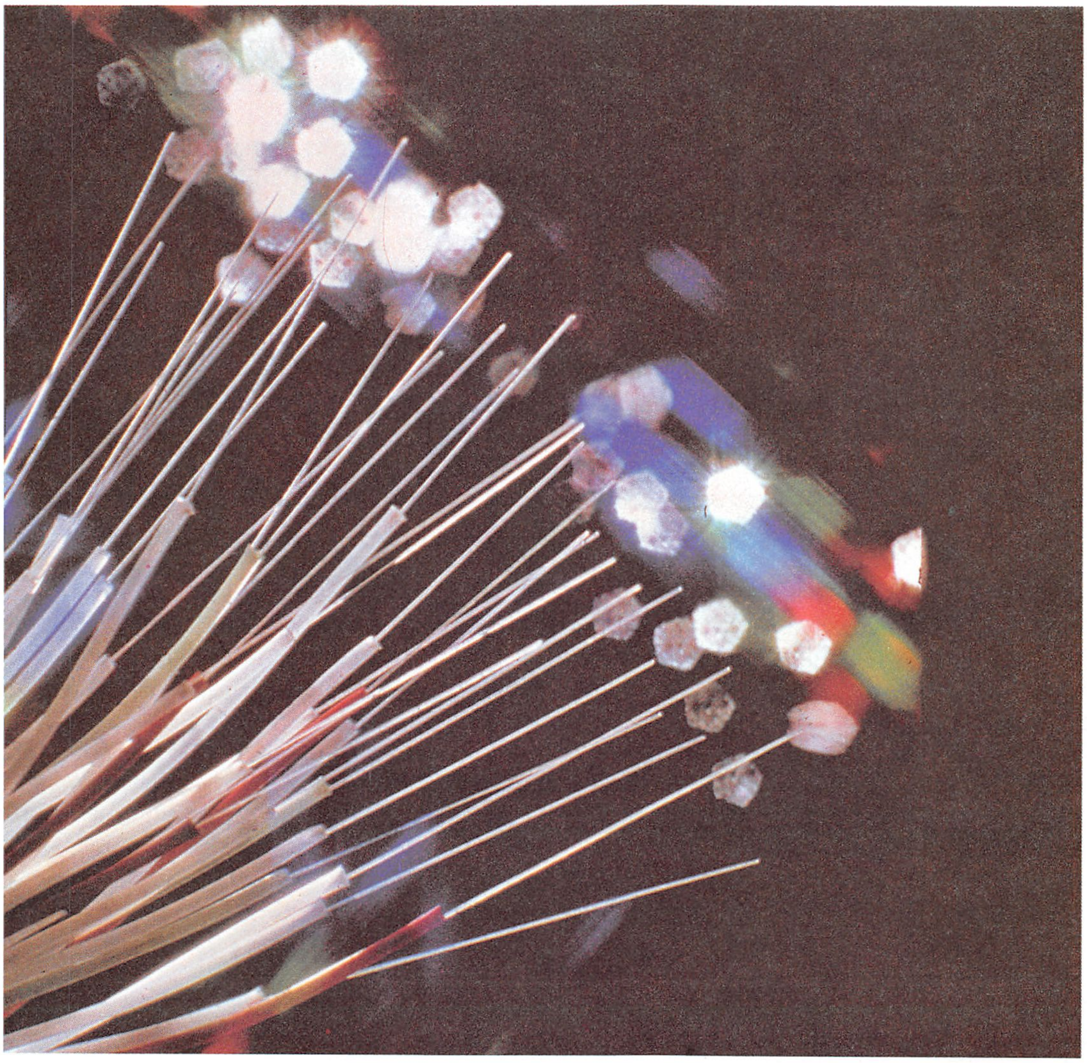
Fibre optic communications systems have

several advantages over cable based systems:
1) Fibre optics are not susceptible to cross-talk and interference, and do not require screening, like electrical cables.
2) Fibre optic cables can be laid almost anywhere, even in sewers, as it is not necessary for them to have the same watertight controlled environment that are necessary for electrical cables.
3) The 'per user, per kilometre' price of optical fibre will fall, sometime in the future, below that of copper cable, thereby reducing costs.
4) The inherent properties of the optical fibre system, mean that long-term maintenance will be reduced as breakdowns will be less frequent than with copper cables.
5) Very high speed data transmission is possible – presently up to about 500 Mbits per second.

All communications systems suffer

19





The Research House/British Telecom

19. Cross-sections of three types of fibre optic strands.

Left: fibre optic cable.

signal attenuation over long distances, and **repeater stations** – which are basically booster amplifiers – need to be inserted into the line at fixed intervals. At present, fibre optic transmission systems require these repeater stations to be placed closer together than they would be for conventional cable systems.

Glossary

aerial	the part of a radio communications system that either radiates or receives electromagnetic energy to and from space
coaxial cable	cable with two conductors, one inside the other, both sharing the same axis. Useful in high frequency applications
dipole radiator	the most basic and important type of aerial. Has two conductors, each a quarter of a wavelength in length, in line. Can be horizontally or vertically polarized
direct wave	radio waves that travel directly to their receiver. Also known as line of sight transmission
electromagetic radiation	a form of energy that travels in waves, radiated from their point of origination. Is composed of both electrical and magnetic transverse fields
electromagnetic spectrum	the range of frequencies of electromagnetic waves, of which radio waves and visible light each form a part
ground wave	radio wave that travels along the ground – also known as surface waves. These are in the long wave band
Marconi aerial	quarter wave aerial that is, in effect, half a dipole earthed at its lower end – the ground reflecting the other half
microwave	electromagnetic radiation with a wavelength in the range 0.003 m to 0.3 m
optical fibre	fine, high quality glass fibre, along which light can be transmitted. Used in modern digital high speed communication systems
radiation pattern	the characteristic pattern of electromagnetic waves propagated by an aerial
sky wave	radio wave that is propagated into the sky and reflected back to earth by the layers of the ionosphere
telecommunications	the branch of applied science that is concerned with the transfer of information by electromagnetic means. Literally translated from the Greek, means communications at a distance
waveguide	tubular or rectangular sectioned metal structure that can be used for the transmission of microwaves. Used either in private links, or as feeders to and from microwave aerials

BASIC THEORY REFRESHER

ELECTRICAL TECHNOLOGY

Network theorems

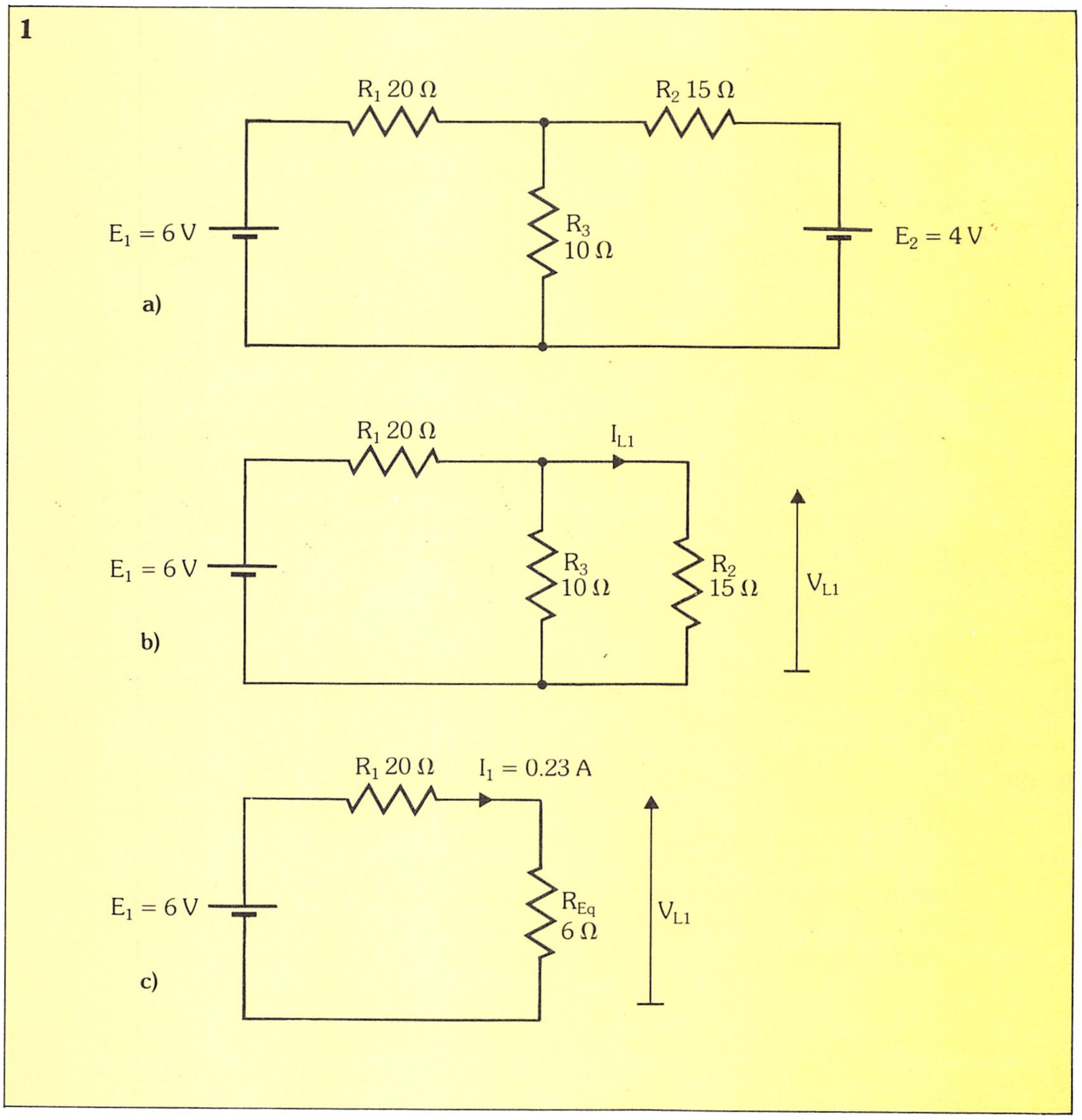
In the previous *Basic Theory Refresher* we studied a general approach to the solution of complex networks based on Kirchhoff's two theorems. Although this approach is perfectly adequate and may be used to solve every electrical network, its application in *certain* networks is rather difficult. To help in the solution of such networks, a number of theorems have been devised which permit simplification of a network into an equivalent circuit before attempting to solve it.

Superposition theorem

One of the most useful of these many theorems is the **superposition theorem**. This states that in all *linear* networks (those composed of resistors, capacitors, inductors, transformers and amplifiers), the total effect of a number of different voltage or current sources is found by adding together the effects of each source acting on its own (with all other voltage sources replaced by short circuits, and all other current sources replaced by open circuits). This allows us to solve a network for *each* generator within it, and then to *add* the results.

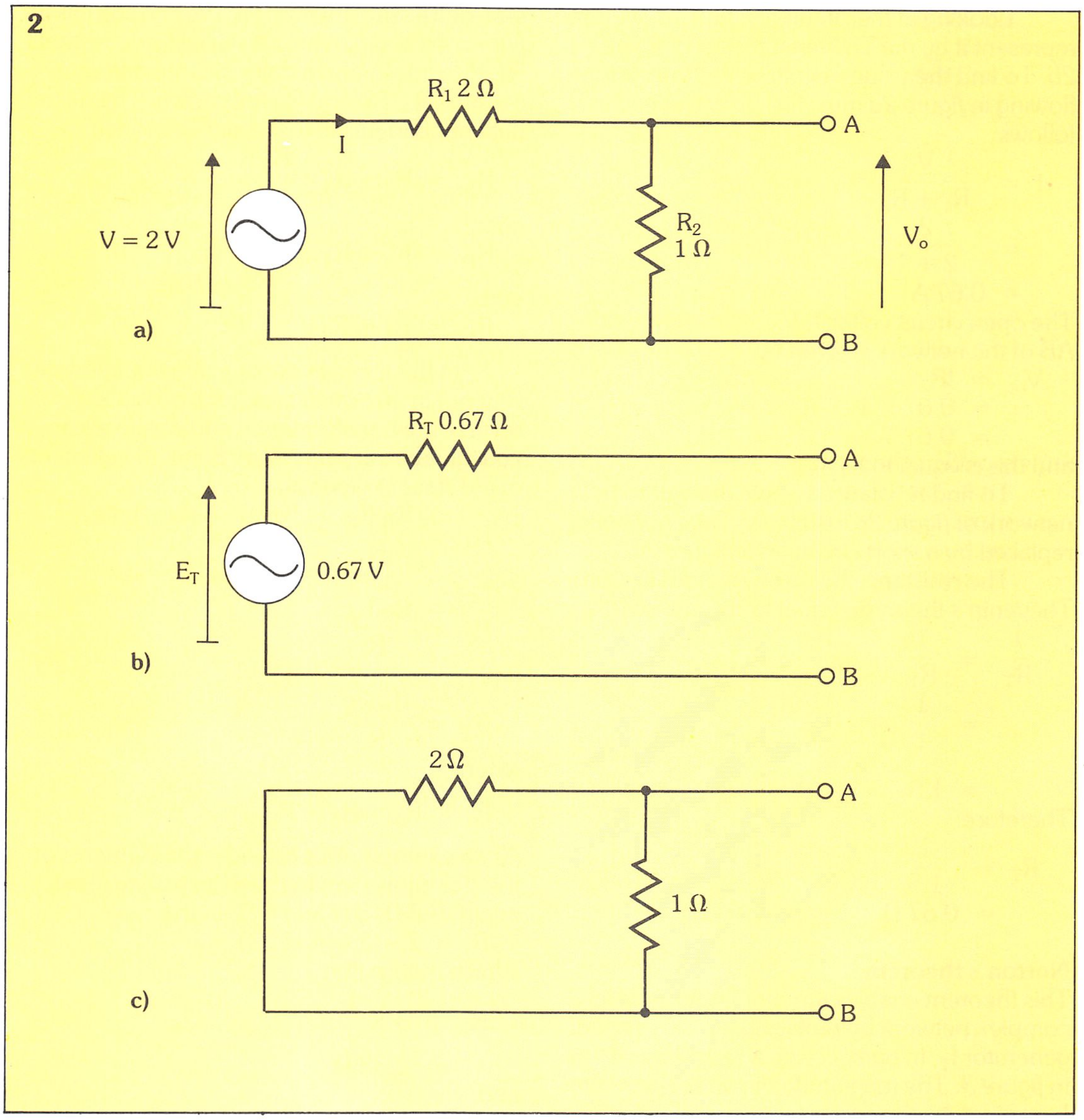
This can be illustrated with the simple example of *figure 1a*, where we wish to find the current flowing in resistor R_3, when two batteries, E_1 of 6 V and E_2 of 4 V, are connected through two resistors, R_1 and R_2.

Let us first consider battery E_1 alone, and draw the circuit as in *figure 1b* with battery E_2 replaced by a short circuit. The two resistors R_2 and R_3 are in parallel and therefore may be



1. Solving a complex linear network using the superposition theory.

2. Using Thévenin's theorem to solve the network shown in (a).



replaced (as in *figure 1c*) by the single resistor R_{Eq} which is given by:

$$\frac{1}{R_{Eq}} = \frac{1}{10} + \frac{1}{15}$$

which gives $R_{Eq} = 6\ \Omega$.

The current, I_1, flowing in this circuit is given by:

$$\frac{6}{20 + 6} = 0.23\text{ A}$$

and the voltage:

$$V_{L1} = 0.23 \times 6 = 1.38\text{ V}$$

Returning to *figure 1b* we see that this voltage will be driving a current I_{L1}, through resistor R_3, which is given by:

$$I_{L1} = \frac{1.38}{10} = 0.138\text{ A}$$

We may continue in a similar manner for the battery E_2, by replacing battery E_1 with a short circuit. In this case, we obtain the current I_{L2}, through resistor R_3, as 0.123 A.

The total current flowing in this resistor, as a result of both the batteries connected together, is:

$$I_L = 0.138 + 0.123 = 0.261\text{ A}$$

Thévenin's theorem

This theorem states that any complex two terminal circuit, containing several voltage generators and a number of resistances, may be replaced by a circuit model consisting of a single voltage generator, E_T, in series with a single resistance, R_T. The value of generator E_T, is the voltage which would be measured between the terminals of the network, if no load is connected – the open circuit voltage. Resistor R_T's value is the resistance which would be measured between the terminals if the voltage generators were replaced by short circuits.

Looking at the circuit of *figure 2a*, we can represent it by the Thévenin model of *figure 2b*. To find the value of voltage E_T, the current flowing in *figure 2a* must first be calculated as follows:

$$\begin{aligned} I &= \frac{V}{R_1 + R_2} \\ &= \frac{2}{2+1} \\ &= 0.67\ \text{A} \end{aligned}$$

The open circuit voltage, V_{oc}, at the terminals AB of the network is given by:

$$\begin{aligned} V_{oc} &= IR_2 \\ &= 0.67 \times 1 \\ &= 0.67\ \text{V} \end{aligned}$$

and this is equal to E_T.

To find resistance R_T, we draw the network of *figure 2a* with the voltage generator replaced by a short circuit, as in *figure 2c*.

The resistance between A and B is, from Thévenin's theorem, equal to R_T, and so:

$$\begin{aligned} \frac{1}{R_T} &= \frac{1}{R_1} + \frac{1}{R_2} \\ &= \frac{1}{1} + \frac{1}{2} \\ &= 1.5 \end{aligned}$$

Therefore:

$$\begin{aligned} R_T &= \frac{1}{1.5} \\ &= 0.67\ \Omega \end{aligned}$$

Norton s theorem

This theorem enables the simplification of any complex network by replacing it with a *current generator* I_T, in *parallel* with a resistance, R_T as in *figure 3*. The magnitude of current generator I_T, is given by the current which would flow in a short circuit connected across the output terminals, and the value of resistance R_T, is that which would be measured looking into the terminals of the network if all voltage generators were replaced by a short circuit, and all current generators were replaced by an open circuit.

Star-mesh transformation

The **star-mesh transformation** is a technique whereby a star of branches radiating from a single node, to *n* other nodes, may be replaced by a network with interconnections between every pair of the *n* nodes and completely eliminating the star point. Every time this transformation is applied, therefore, the number of nodes in a network is reduced by one.

The simpliest form the transformation takes is the conversion of a star, having four nodes, into a delta of three branches connecting three nodes – this is known as the **star-delta transformation**. We can see this in *figures 4a* and *b*, where a star of three resistors R_1, R_2 and R_3, and a delta of three different resistors R_A, R_B and R_C are shown. The delta network is identical to the star network if:

$$R_A = R_2 + R_3 + \frac{R_2 R_3}{R_1}$$

and:

$$R_B = R_3 + R_1 + \frac{R_3 R_1}{R_2}$$

and:

$$R_C = R_1 + R_2 + \frac{R_1 R_2}{R_3}$$

In this simple case of a three pointed star (but not in any other case), it is possible to reverse the transformation and obtain a star equivalent to a given delta. Here, the elements of the star are given by:

$$R_1 = \frac{R_B R_C}{R}$$

and:

$$R_2 = \frac{R_C R_A}{R}$$

and:

$$R_3 = \frac{R_A R_B}{R}$$

where:

$$R = R_A + R_B + R_C$$

As an example, let's consider the elements of the star equivalent to the delta having resistances of 2 Ω, 3 Ω and 4 Ω. In this case:

$$R = 2 + 3 + 4 = 9\ \Omega$$

Which means that:

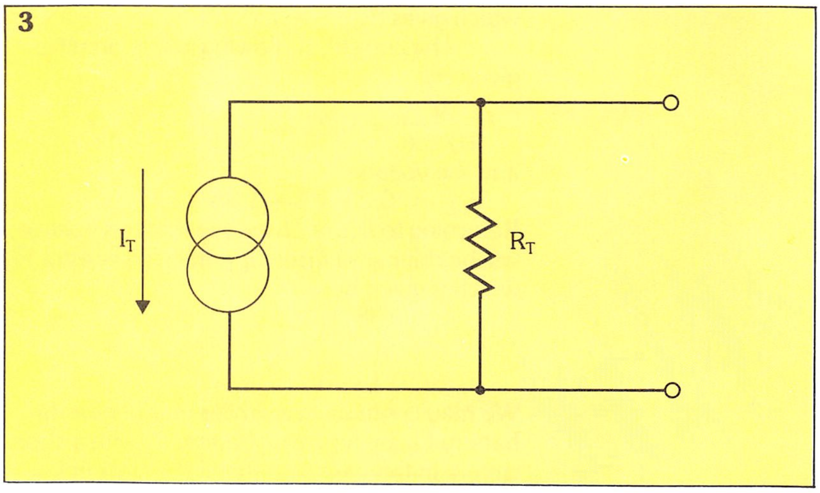
$$\begin{aligned} R_1 &= 3 \times \frac{4}{9} \\ &= 1.33\ \Omega \end{aligned}$$

and:

$$\begin{aligned} R_2 &= 4 \times \frac{2}{9} \\ &= 0.89\ \Omega \end{aligned}$$

and:

3



3. Norton's theorem simplifies any complex network by replacing it with a current generator in parallel with a resistance.

$$R_3 = 2 \times \frac{3}{9}$$
$$= 0.67\,\Omega$$

Maximum power transfer theorem

If we have some network which is represented by its Thévenin model, for example the network shown in *figure 5* connected to a load resistor R_L, then power will be dissipated in the load resistor equal to the product of the current and the voltage. If resistor R is varied from zero to very large values, the power dissipated will also vary. When resistor R_L is very small, the voltage across it will also be small and so, too, will the power dissipated. Again, when resistor R_L is very large, the *current* will be small and so again the power dissipated will be small.

Between the two extremes, however, the power dissipated rises until it reaches its maximum value, when resistor R_L equals the source resistor, R. This is the condition for **maximum power transfer** from the source to the load, and the load is then said to be **matched** to the source.

The value of this maximum power is given by:

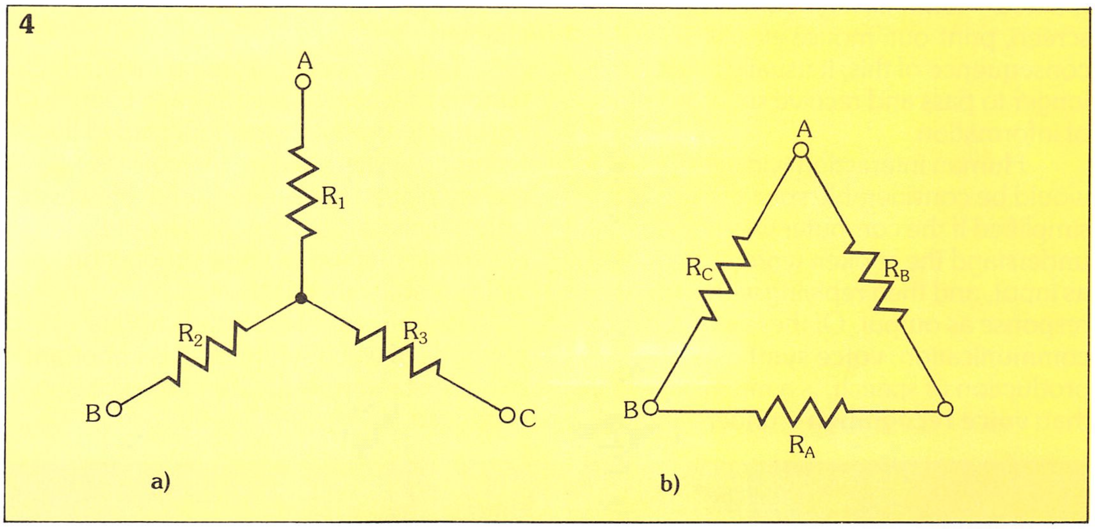
$$P_{max} = \frac{E^2}{4R}$$

and this is termed the **available power of the source**.

Summary

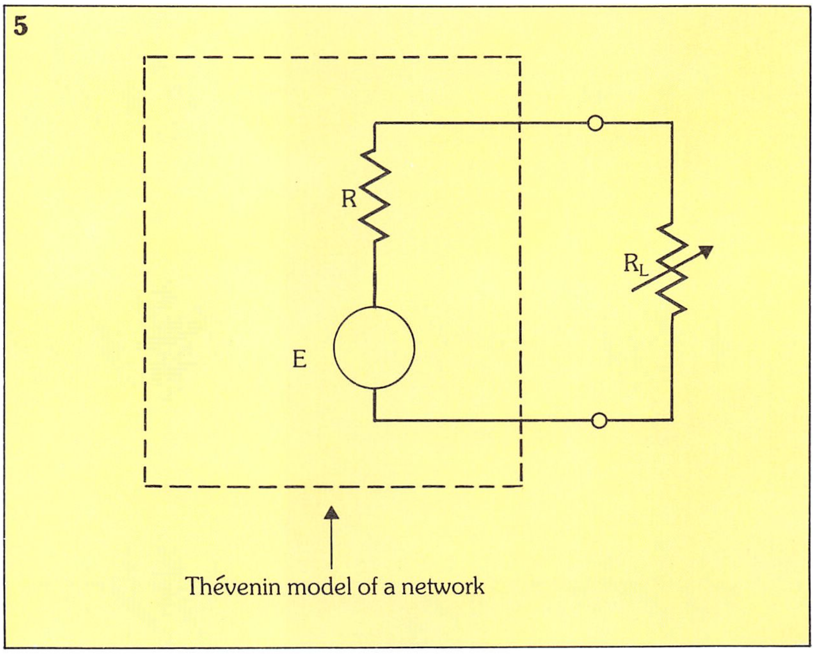
We have seen that all complex networks may be simplified, and equivalent circuits may be drawn of those networks, using a number of network theorems.

Although, the networks we have shown are all DC resistor networks, the theorems may all be used in AC networks with impedances, following the usual rules. □



4. Star-delta transformation.

5. Maximum power transfer theorem.



24

DIGITAL ELECTRONICS

Voice synthesis

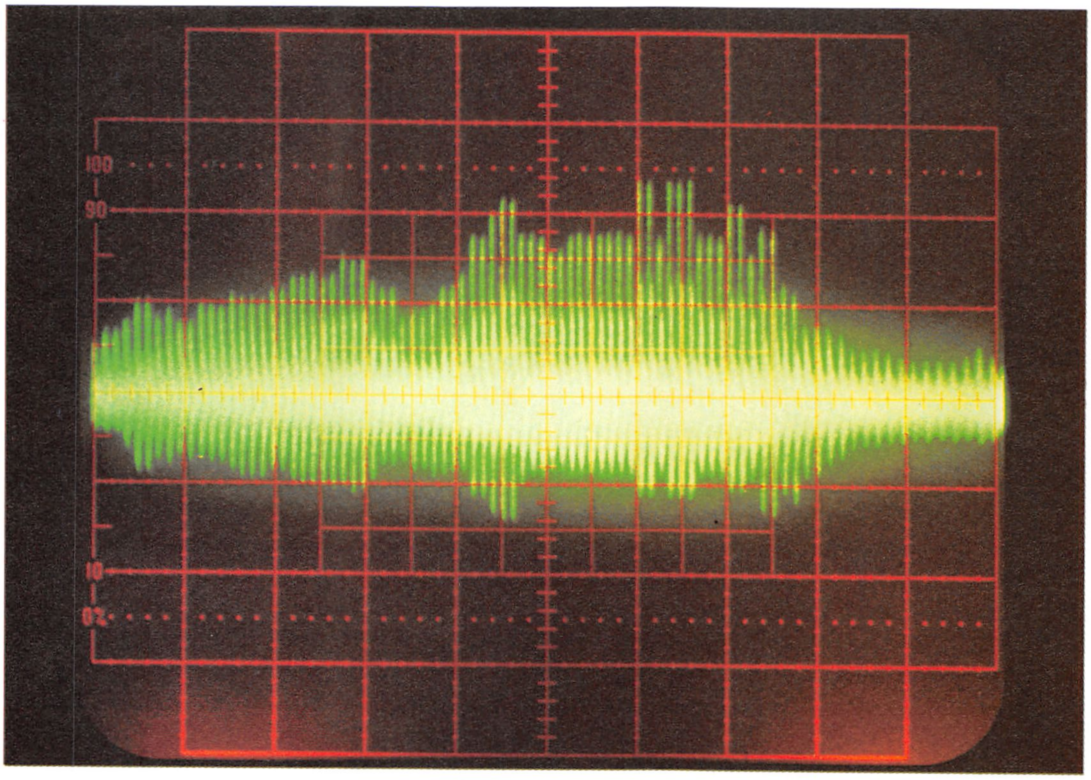
Talking machines

Speech is one of the most effective forms of human communication, enabling direct and rapid two-way transfer of information. Communication with a computer, on the other hand, is usually via an indirect method – terminal keyboard, touch-screen, print-out, mouse etc. As a consequence of this, it usually takes much longer to pass and receive small amounts of information.

Human interaction with computers would be considerably speeded up and simplified if the computer was able to understand the spoken word, accepting it as input, and then replying with a vocal response as output. Of these two areas of communication, **voice synthesis**, the production of speech, is a much easier task than **voice recognition** – understanding the spoken word. Computerised voice recognition is, in fact, a very complex task because each person speaks in a different way – within a single language there are also many different regional accents and different slang phrases. In this chapter we will concentrate on voice synthesis; voice recognition will be discussed in a later chapter.

Talking machines are not new, various mechanical means have been constructed, which were superseded by recorded systems where the voice is merely played back as required. However, only a digital system enables a purely *electronic* method of voice production, which is software controlled.

Before we consider the various methods of voice synthesis, it is important to understand how the human voice is produced.



Paul Brierley

Left: CRT trace of a voice pulse.

Sound

Sound is carried through air as a series of minute compressions and rarefactions of air molecules – a simple way of illustrating this is shown in *figure 1*.

However, this representation does not enable *complex* sound waves to be visualised – to do this, simple waveforms, such as sine waves or square waves, are used to represent the changes in pressure that occur in the air as we speak (*figure 2*).

Sound can be described in terms of its four parameters – pitch, amplitude, timbre and envelope – and may be regarded as a simultaneous combination of these four parameters which change on a continuous basis.

We will now go on to examine each of these parameters in detail.

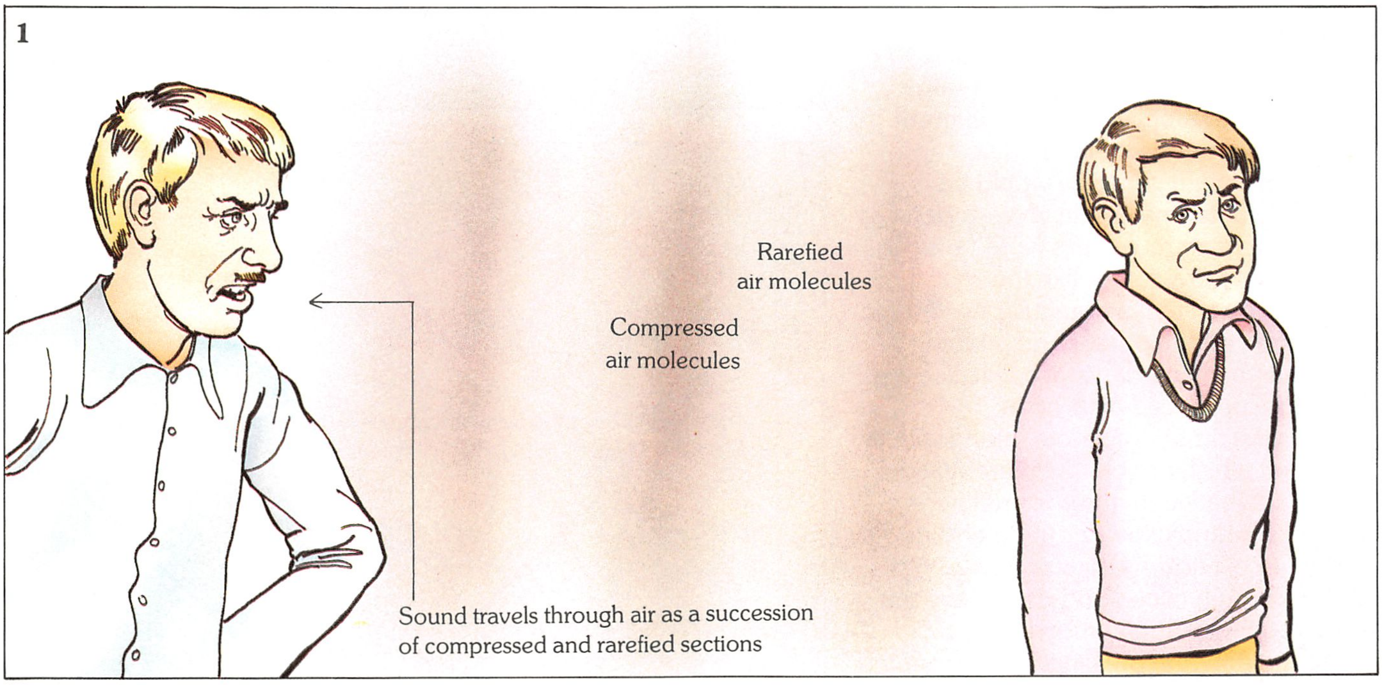
Pitch

The actual tone or frequency of the sound, how high or low the sound is, is termed the **pitch**. For example a soprano singing a single note produces a sound which is higher in pitch than a bass note. The pitch of a sound, in fact, depends on the number of compressions and rarefactions of air molecules detected by a listener over a set time interval. We would say, for example, that a high pitch note, which may comprise 3,000 individual compressions and rarefactions of air molecules during every second, has a frequency of 3 kHz. Looking again at *figure 2*, the two sine waves of different frequencies represent the way in which air molecules are compressed and rarefied.

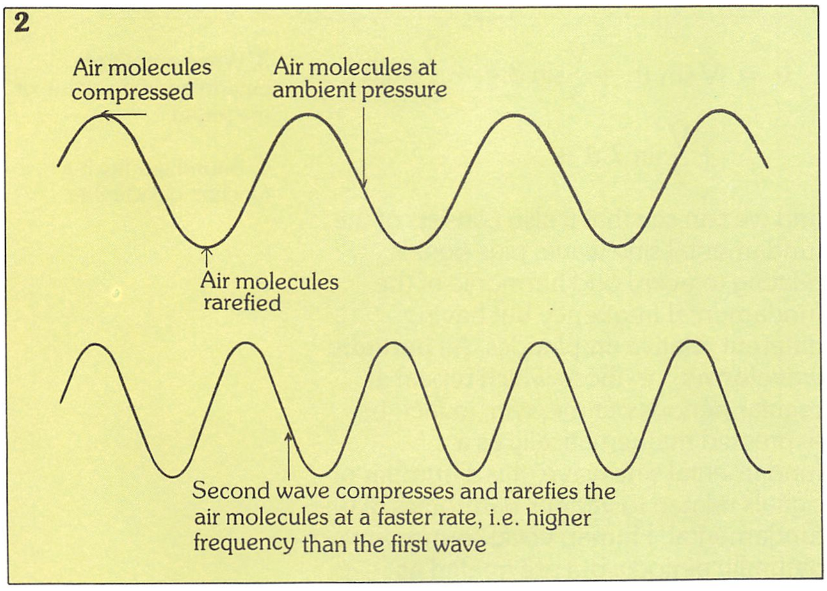
1. Changes in air pressure carry sound.

2. The frequency of a sound determines the pitch – how high or low the sound is.

1


Rarefied air molecules
Compressed air molecules
Sound travels through air as a succession of compressed and rarefied sections

2


Air molecules compressed
Air molecules at ambient pressure
Air molecules rarefied
Second wave compresses and rarefies the air molecules at a faster rate, i.e. higher frequency than the first wave

Amplitude

The **amplitude** of a sound wave represents the volume of the sound; a sound of, say, 3 kHz could be so quiet that it is only just audible, or it may be so loud that it hurts, or even damages the ear, depending on the amplitude of the wave. In terms of the transmission of the sound wave through air, the amplitude is defined by the ratio of the numbers of air molecules in compressed regions to the numbers of air molecules in rarefied regions: for quiet sounds the difference is small; for loud sounds the difference is large. The difference in amplitude of two sine waves, representing sound waves of the same frequency, is shown in *figure 3*.

Timbre

The **timbre** of a sound wave is a function of the overall quality of the sound. This can best be shown by comparing two different waveforms, as in *figure 4*. Here, a sine wave and a triangular wave of the same frequency and amplitude, have different shapes. The sounds that these two waveforms might generate would be noticeably different. A sine wave-type movement of air molecules would sound smooth and pure, rather like the sound produced by a flute; a triangular wave-type movement, on the other hand, would sound harsh in :omparison, a little like the same note proc iced by a violin.

It is the timbre of a note (along with its envelope) which enables a differentiation to be made between notes from different instruments, and allows us to determine what source a non-musical sound comes from. Although two sounds may have the same frequency, their timbres may differ due to the differing presence of **harmonics** of the fundamental frequency. For example, a sine wave is pure, with no harmonics, and may thus be represented mathematically by the fundamental expression:

$$v = \hat{V} \sin \theta$$

A typical triangular wave, on the other hand, has a more complex representation:

$$v = \hat{V} \sin \theta - \frac{\hat{V}}{9} \sin 3\theta + \frac{\hat{V}}{25} \sin 5\theta - \frac{\hat{V}}{49} \sin 7\theta + \ldots.$$

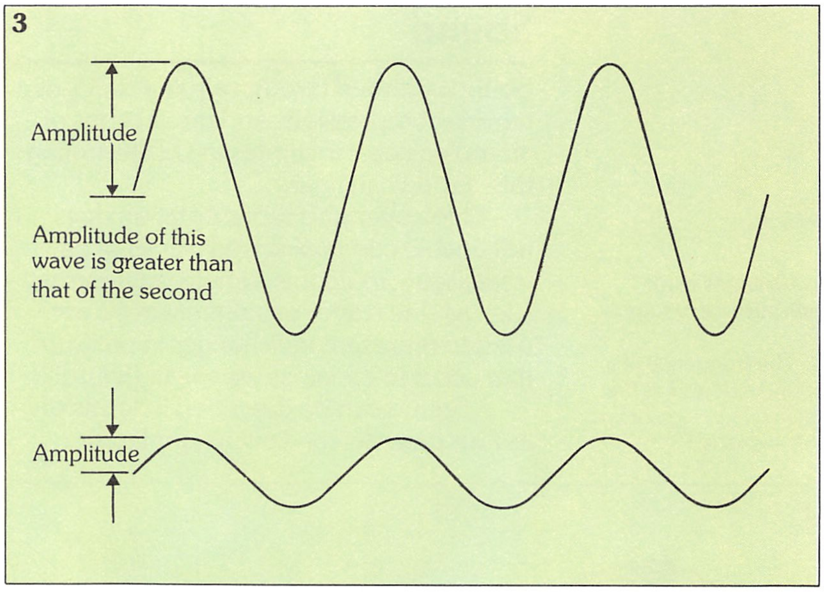
That is, it contains signals related to every odd harmonic of the fundamental.

Similarly, a square wave can be represented mathematically by the expression:

$$v = \hat{V} \sin \theta + \frac{\hat{V}}{3} \sin 3\theta + \frac{\hat{V}}{5} \sin 5\theta + \frac{\hat{V}}{7} \sin 7\theta + \ldots.$$

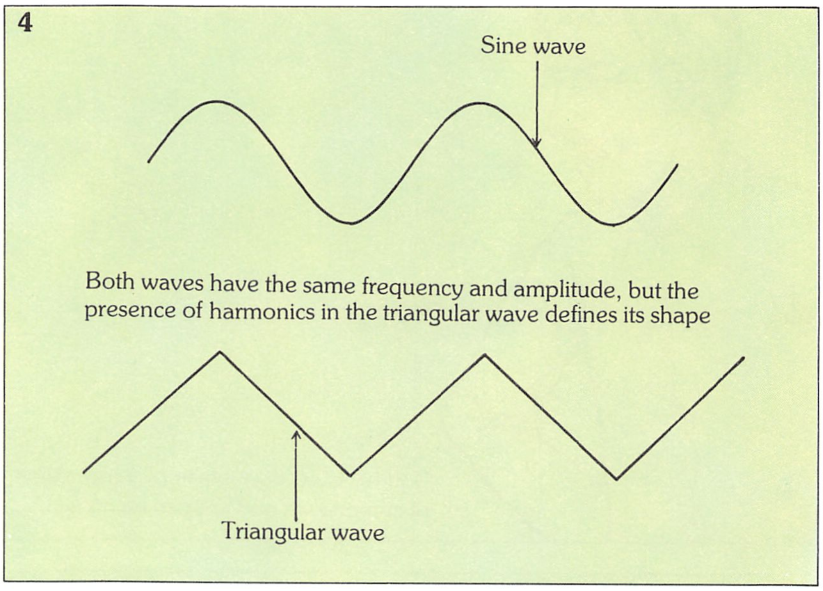
and we can see that it also consists of the fundamental sine wave, plus signals relating to every odd harmonic of the fundamental frequency but having different relative amplitudes. All **periodic waveforms**, i.e. those which repeat at regular periods of time, can, in fact, be expressed mathematically as a fundamental sine wave plus a number of signals related to various harmonics of the fundamental. Human vocal sounds are not naturally periodic but are treated as

3



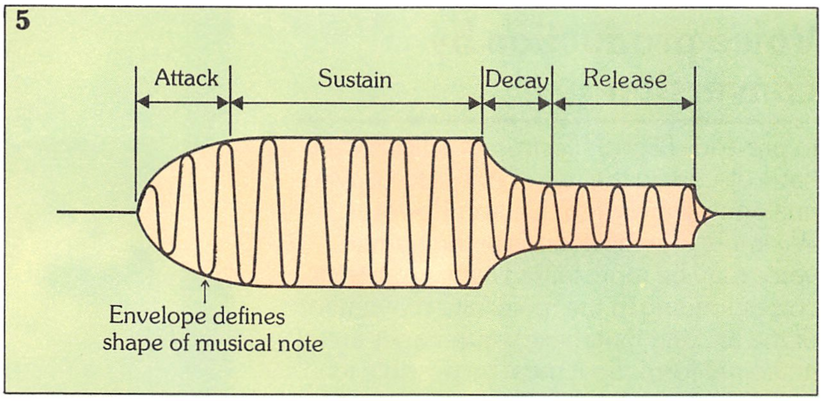
3. Wave amplitude determines the volume of the sound.

4


Sine wave
Both waves have the same frequency and amplitude, but the presence of harmonics in the triangular wave defines its shape
Triangular wave

4. Sound quality is a function of its timbre.

5



5. The overall shape of the sound wave is known as its envelope.

pseudo-periodic.

Envelope

The overall shape of a sound wave is known as its **envelope**. *Figure 5* shows a sine wave of a fixed frequency whose amplitude is defined by a particular envelope. In musical terms, envelopes of any notes may generally be split up, as shown in *figure 5*, into four basic time intervals: **attack**, which is the time taken for the sound to build up from zero to maximum amplitude; **sustain**, during which time the maximum amplitude is maintained; **decay**, where the amplitude falls to a small value; and **release**, during which time the smaller amplitude is held until finally the note ends.

Non-musical sounds, on the other hand, may have much more complex envelopes.

The voice

The human voice, of course, consists of many sounds, both musical and non-musical, and the way in which it is produced in speech or song has relevance to the ways it can be digitally produced.

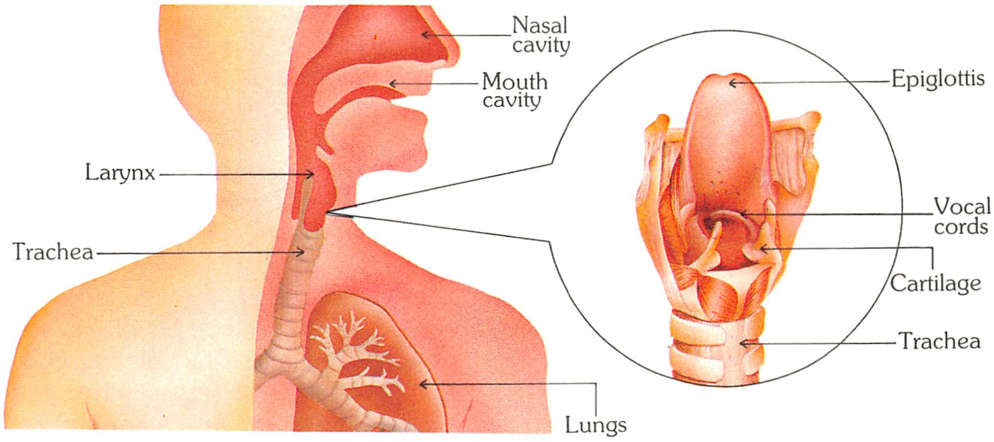
Human vocal reproduction is a complex combination of muscular activities involving the trachea or windpipe, the vocal cords, larynx (voice box within the throat which holds the vocal cords), and the mouth and nasal cavities as shown in *figure 6*.

The vocal cords consist of a pair of membranes attached to the walls of the larynx. Their appearance and action is somewhat like a diaphragm with a slit. As the membrane tissue is elastic, air forced up through the slit via the lungs and trachea causes the membrane to vibrate. This vibration of the vocal cords alternately blocks and releases air up into the mouth and nasal cavities, producing the compressions and rarefactions of air molecules which we know as sound waves.

Muscles in the larynx control the tension of the vocal cord tissue, determining the frequency of vibration of the vocal cords. Thus, the pitch of the sound produced may be varied by muscular movement, and is usually within the range of about 100 Hz to 4 kHz.

The volume of the sound produced

6



6. Human voice production involves many complex muscular activities.

depends on the amount of air forced through the vocal cords in a given time; this, in turn, depends on the rate at which the lungs exhale.

The sound, at this stage, will be very rich in harmonics, and its timbre is controlled by the positioning and shape of the mouth and nasal cavities, along with the lips, teeth and tongue, introducing further harmonics, or reducing those already present.

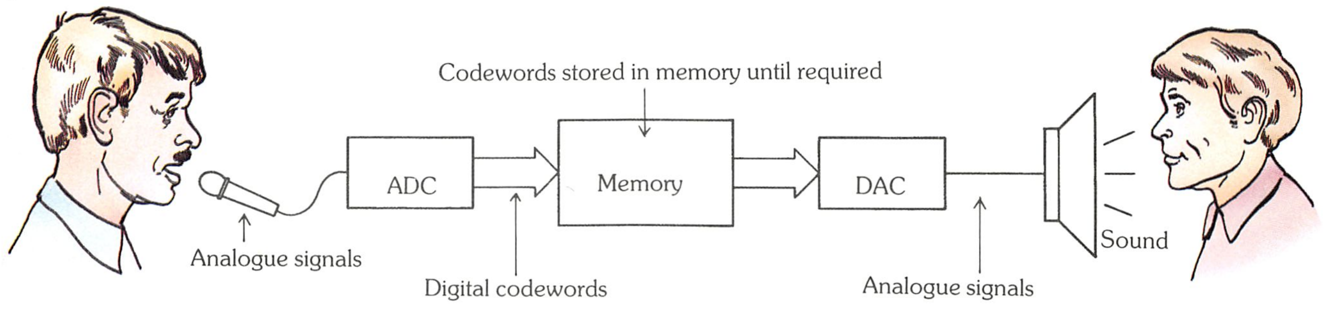
Lips, teeth, tongue, cavities, and amount of air through the vocal cords may all combine to produce many envelopes of vocal sounds.

Even with this brief description of human voice production, we can see that any digital representation of voice will be necessarily complex. We will now examine the three main electronic methods in use in voice output systems.

Voice production by conversion

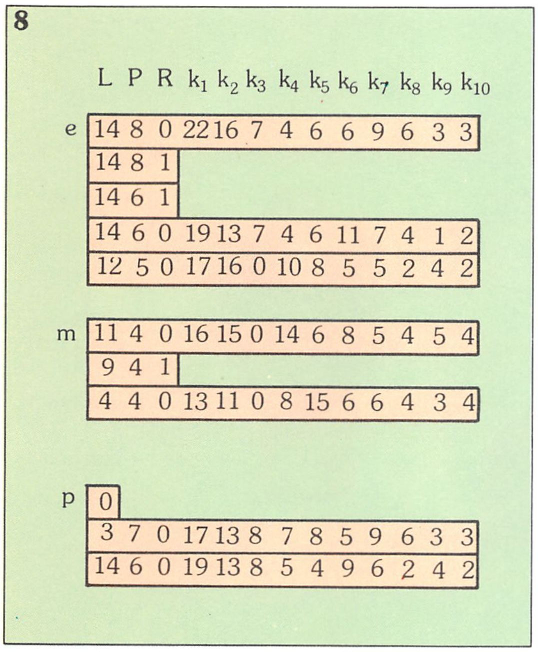
In previous *Digital Electronics* articles we have discussed the use of ADCs to convert analogue signals into digital codewords. We know that the sounds of the human voice may be represented by waveshapes, corresponding to the molecular movement of the air, and that these waveshapes are of *analogue* form. So it must be possible to convert this analogue signal into a number of *digital codewords*, and then write these codewords into a digital memory. At some later time, these codewords may then be read from memory, converted back to analogue form and reproduced as sound corresponding to the original voice. The process is shown in *figure 7*, where a microphone is used as a transducer to convert sound waves into an analogue electrical signal. The analogue signal is then converted by an ADC to digital codewords which are stored in memory. It is then a matter of accessing the required memory locations and reading the consecutive codewords which make the required sections of voice. After digital-to-analogue conversion, the analogue electrical signal may finally be amplified and converted by the loudspeaker into sound.

7



7. Converting analogue sound waves into digital codwords to be stored in memory; then reconverting them back into sound.

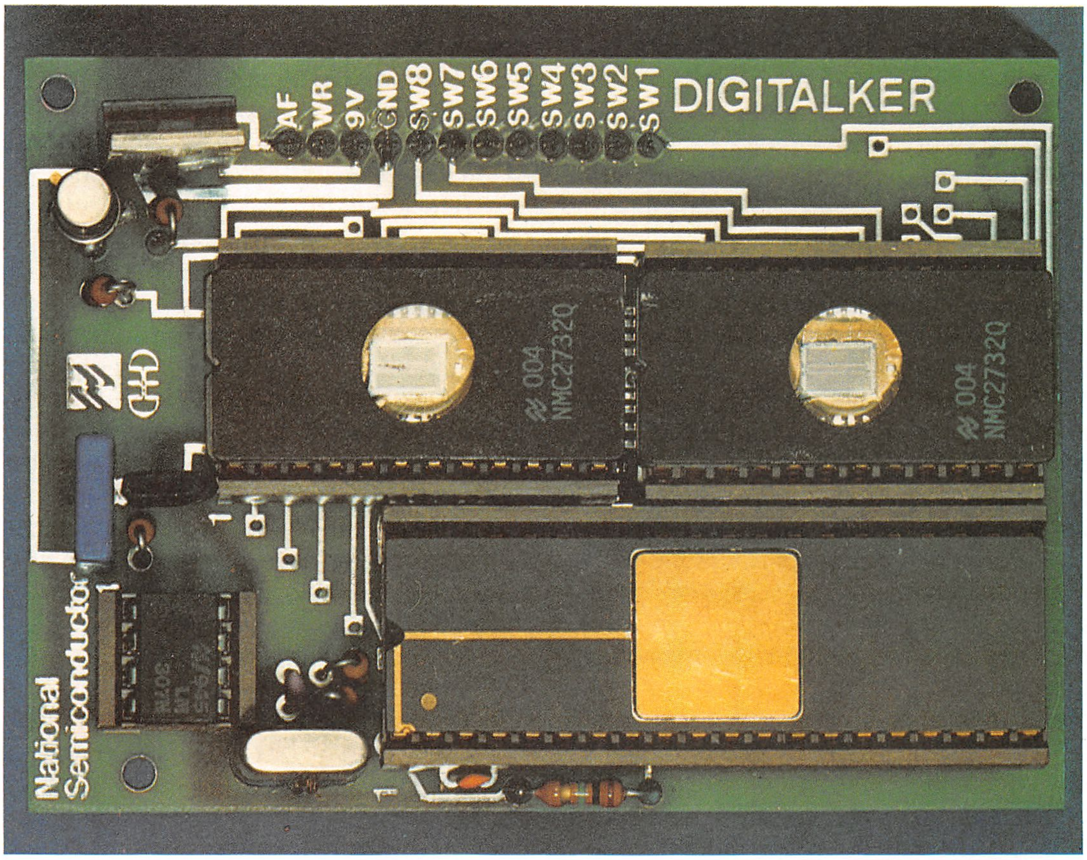
8



	L	P	R	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}
e	14	8	0	22	16	7	4	6	6	9	6	3	3
	14	8	1										
	14	6	1										
	14	6	0	19	13	7	4	6	11	7	4	1	2
	12	5	0	17	16	0	10	8	5	5	2	4	2
m	11	4	0	16	15	0	14	6	8	5	4	5	4
	9	4	1										
	4	4	0	13	11	0	8	15	6	6	4	3	4
p	0												
	3	7	0	17	13	8	7	8	5	9	6	3	3
	14	6	0	19	13	8	5	4	9	6	2	4	2

8. Using redundancy in the LPC method of voice synthesis.

The block diagram of *figure 7* shows very well that this type of voice production system is not actually a voice *synthesis* system (by the true sense of the word), because the vocal output is merely a recreation of the vocal input. Even though the hardware is completely portable and any of the stored voice messages may be recreated at any time, no synthesis of voice takes place – we can think of the system as a solid state tape recorder, from which *fixed* voice messages may be replayed as and when required.



Right: DIGITALKER speech synthesis system comprising multiple n-channel MOS ICs. When connected to an external filter, amplifier and speaker it synthesises human speech. (Photo: National Semiconductor).

This type of system samples the analogue signal into a series of *pulses* at a succession of instants of time. Each of these pulses is coded by a digital codeword. It is thus known as **pulse code modulated** (PCM) voice output.

Such a voice production system works extremely well but the principle suffers one major drawback – the amount of memory required. This can be seen if we consider the following example of a speech message, of say, 10 seconds in length. We know that sampling of a signal must take place at a rate of at least approximately twice the maximum frequency, in order that aliasing does not occur (see *Digital Electronics 20*), so the sampling rate of an analogue signal corresponding to voice (in the range of approximately 100 Hz to 4 kHz) must be about 8000 samples per second.

Now, a 10 second speech message must therefore be sampled 80,000 times. Each sample is converted to a digital codeword, of which the word length affects the final voice quality – this is because the more bits used in each codeword, the lower the quantization error. Generally speaking, for reasonable quality reproduction of voice, a quantization error of less than about ± 0.2% is required (in comparison, a quantization error of about ± 0.01% is required for digital reproduction of high quality music), which corresponds to the use of 8-bit codewords. Each of the 80,000 samples of the 10 second message therefore needs eight bits of storage, meaning that a total of 640,000 bits of storage are required. That is, *over half a megabyte of memory*!

It is obvious, therefore, that even for a small collection of different messages, an enormous quantity of memory is required. This memory requirement is the main restriction of the use of this type of voice production method in electronic voice output systems.

Linear predictive coding

The PCM method of voice production relies on storing data corresponding to only one parameter of the speech signal (i.e. its actual analogue value) 8,000 times every second. In this respect, none of the other parameters such as pitch, amplitude, timbre, or envelope need to be defined and stored – they are automatically defined by the 8,000 times per second sample. However, these parameters of a speech signal change at a far lower rate than this (because of the finite time taken by the body to move all the muscles, membranes, tissues and bones involved in speech production). In fact, any speech signal can be broken down into time slots of about 20 to 25 ms duration.

One voice synthesis method, **linear predictive coding**, relies on this principle, breaking down the speech signal at each time slot into a number of parameters which may be digitally coded and stored.

Table 1 gives a list of possible parameters which may be measured in such a system, along with the numbers of bits required to define each one. Only the first three parameters have been defined – others may be, say, loudness of harmonic frequencies etc. The thirteen parameters shown may not be sufficient to allow high quality speech reproduction so more practical systems may use many more parameters. A total of 48 bits is needed, in our example, to adequately define thirteen parameters.

Using this system as it stands, the 10 second speech signal we used in our previous example would require:

$$10 \times 48 \times 50 = 24{,}000 \text{ bits}$$

– already a significant reduction in the amount of memory required to store the signal.

Still further reductions in the size of memory can be made if the method takes into account the large amount of **redundancy** which occurs in such a speech model. For example, *figure 8* shows three letters together with possible parameter values, which may be obtained in such a system, when they are spoken. The letter 'e' has been broken up into five time slots (known as **segments** or **frames**). The first segment defines the *eee* type sound made when saying the letter, but as this is pronounced over three segments, the second and third segments are merely repeats of the first. This is signified by the third parameter: a repeat bit. When this repeat bit is at logic 1, none of the following parameters need be defined. Loudness and pitch, however, before the repeat bit, may be changed. The fourth and fifth segments are defined by a full list of parameters to make up the whole sound of the letter.

The letter 'm' is seen to consist of three segments, the first being repeated as the second. Finally, the letter 'p' consists of three segments, the first of which has zero loudness, corresponding to the time taken to move the lips into the correct position before voicing the letter. As the loudness is zero, no further parameters need be present. Pauses between words and sentences will also be in this form. As many of the segments are thus expressed as a combination of preceding ones, the principle of this method is known as linear predictive coding (LPC).

Obviously, the reduction of the number of parameters in those segments affected by redundancy, lowers the overall number of storage bits required. Average numbers, for a 10 second speech signal, are around 10,000 bits – a reduction of over 50 to 1 from the pulse code modulated method.

Table 1
Possible parameters used in the LCP system

Parameter	Maximum number of values	Number of bits required
Loudness	15	4
Pitch	32	5
Repeat	2	1
k_1	32	5
k_2	32	5
k_3	16	4
k_4	16	4
k_5	16	4
k_6	16	4
k_7	16	4
k_8	8	3
k_9	8	3
k_{10}	4	2

Total number of required bits = 48

Voice synthesis

LPC methods of voice production can be very effective as the basis behind voice output systems. For example, modern cars which have a 'talking computer' operate on the LPC principle. Generally, a microprocessor IC is used to control output from a dedicated ROM device, mask programmed with a selection of chosen words. The words within a ROM may be quite specific for a given application, and changing the ROM to another with different words allows the system to be used in a different application.

Although such LPC systems are very versatile, the range of words within an application still depends on the memory size, and a speech system for a specific application will rely on there being a suitable ROM available. Both of these factors exist because such an LPC system is, like a PCM system, not an actual synthesis system – it can only output what is previously input.

The final type of system we will now consider is, however, a voice *synthesis* system. It relies on the fact that speech can be broken down, not into 8,000 samples per second of one parameter, or 50 samples per second of many parameters, but instead into a relatively small number of individual sounds which combine together to make words. The individual sound-types are known as **phonemes** and consist of sounds formed by vowels, resonants, fricatives, stops etc. Within each phoneme-type there may be a number of slightly different pronunciations known as **allophones**.

A collection of allophones (enough to be able to accurately reproduce all the sounds used in a language – say 50 to 150, depending on the required speech quality) are stored in ROM. It is then a straightforward task for a software controlled microprocessor to call up the sounds in the sequence necessary to produce *any* word or phrase. Voice production is thus truly synthetic, and a large memory size is not required. Memory size is also not dependent on the length of the speech signal – only on the number of allophones stored – and is thus fixed.

Glossary

allophone	one of two or more forms of the same phoneme
envelope	overall loudness parameters of a sound. A note from a musical instrument has an envelope which can be approximated by four individual parts – attack, sustain, decay and retain
linear predictive coding (LPC)	principle, used in electronic voice production, in which a speech signal is broken down into time slots (segments or frames) specified by a number of parameters
periodic waveforms	waveforms which may be mathematically represented as a fundamental sine frequency plus a number of harmonics of that frequency
phonemes	individual sound types within a language
pulse code modulation (PCM)	principle, used in electronic voice production systems, where a speech signal is sampled, the analogue value of the signal at each sample is converted to digital codewords, and the codewords are stored in a memory
segment, frame	time slots in a linear predictive coding system
timbre	overall quality of a sound

26

SOLID STATE
ELECTRONICS

Analogue circuits-1

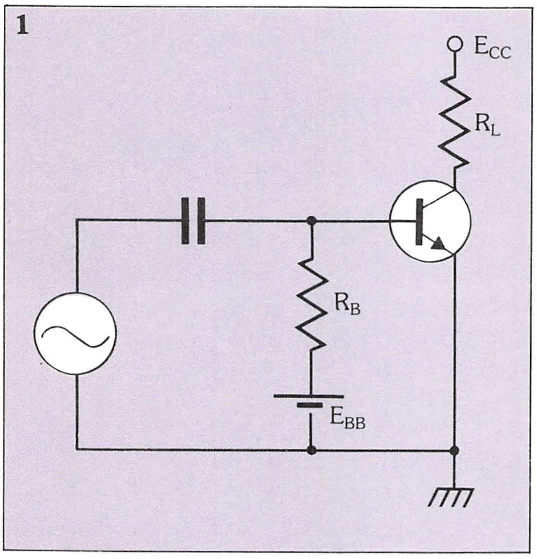
Practical amplifiers

So far we have looked at the way that amplifier circuits operate with small variations in their input signal voltages, giving amplified, but still quite small output signal voltages. However, a great many applications demand that the input voltage is raised to a value high enough to operate a connected device. Consider an ordinary record player. The input to the amplifier from the stylus and cartridge may be only a fraction of a millivolt, while the output voltage required may have to be in the range of several volts – a gain of, say, 5000. Such an application also needs a constant voltage gain over the complete range of musical frequencies.

The gain required from an amplifier is invariably much greater than can be obtained from a single transistor amplifier stage, so we must **cascade** two or more amplifiers together. If we cascade two identical amplifiers of voltage gain, K, and bandwidth, f, we, in fact, produce an amplifier with a gain of K^2. However, we also find that the bandwidth has been reduced to 0.643f! If we cascaded *three* amplifiers together, to obtain a gain of K^3, then the bandwidth would be reduced to 0.51f! This means that if we are designing an audio amplifier that is to have a bandwidth of 20 kHz, and we find that three amplifier stages are necessary to give the gain required, each stage needs to have a bandwidth of:

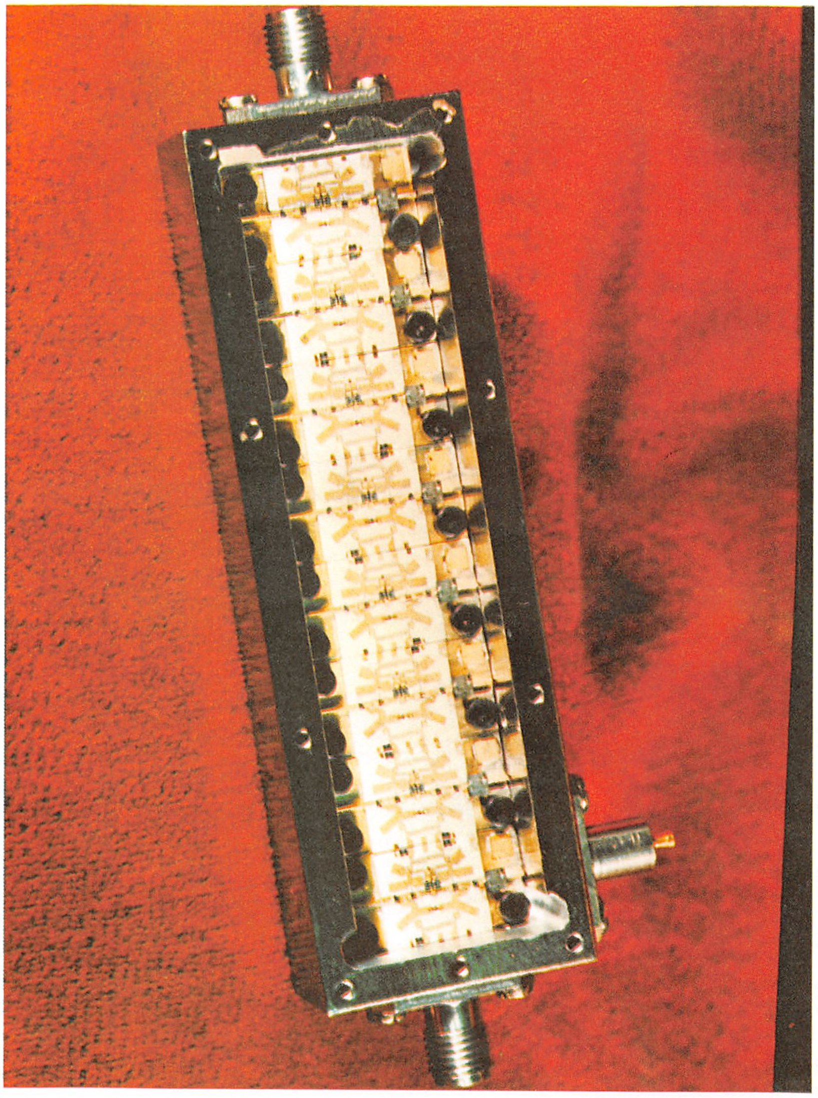
$$\frac{20}{0.51} = 39.2 \text{ kHz}$$

1



1. Simple transistor amplifier.

Below: transistor amplifier.

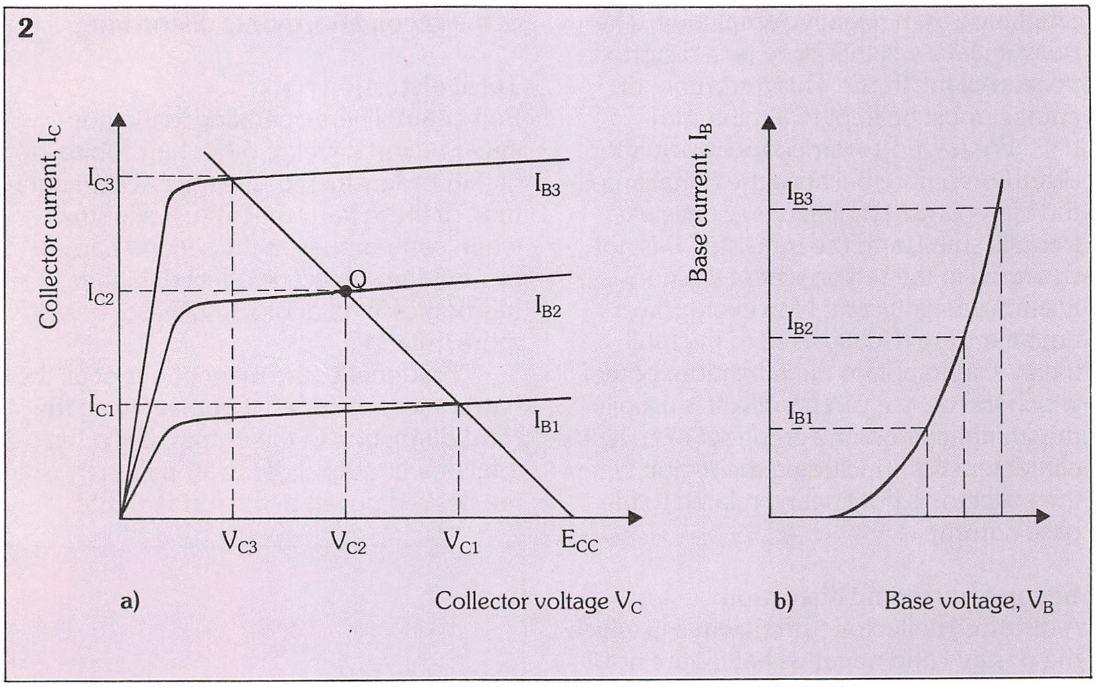


The Research House/Marconi Electronic Devices Ltd

Power amplifiers

The amplifier circuits that we have examined so far have been designed to amplify voltage. However, the final stage of most systems needs to supply *power* to some mechanical device – such as a loudspeaker. The simple transistor amplifier in *figure 1*, for example, can be used to supply power to a load, represented by the resistance R_L. The battery E_{BB} and resistor R_B are used to bias the transistor at a point in the middle of its operating range (point Q, in *figure 2a*, where the base current is I_{B2}). If the base current is then varied sinusoidally about I_{B2}, from a maximum value of I_{B3} to a minimum value of I_{B1}, then the collector current will, as we know, also be sinusoidal, varying from a maximum of I_{C3} to a minimum of I_{C1}. The collector voltage will, of course, also fluctuate similarly – between V_{C3} and V_{C1} shown here.

2. Transfer characteristics for the amplifier in *figure 1*.



The maximum amplitude of the collector sinusoidal voltage, $\hat{V}_C$, is given by:

$$\hat{V}_C = \frac{V_{C3} - V_{C1}}{2}$$

and the amplitude of the collector current, $\hat{I}_C$, is given by:

$$\hat{I}_C = \frac{I_{C3} - I_{C1}}{2}$$

While the rms value of the voltage, V_C, and current, I_C, is given by:

$$V_C = \frac{\hat{V}_C}{\sqrt{2}}$$

$$I_C = \frac{\hat{I}_C}{\sqrt{2}}$$

The power, P_L, dissipated in the load resistor, R_L, is:

$$P_L = V_C I_C = I_C^2 R_L$$

which can be directly expressed as:

$$P_L = \frac{(V_{C3} - V_{C1})}{2\sqrt{2}} \frac{(I_{C3} - I_{C1})}{2\sqrt{2}} = \frac{(V_{C3} - V_{C1})(I_{C3} - I_{C1})}{8}$$

This expression shows us that the power in a load can either be obtained directly, by determining the maximum and minimum current and voltage from the load line drawn on the transistor characteristics; or, if we are working with a practical circuit, these values can be measured using test equipment.

Distortion

We have previously assumed that the collector characteristics consisted of approximately parallel, equidistantly spaced lines. As a consequence of this, a sinusoidal base current, in which the maximum positive magnitude is equal to the maximum negative magnitude, will give rise to a collector current and voltage which will also have identical positive and negative magnitudes. The output signal will thus have the same shape as the input signal and there will be no distortion.

However, the characteristics of a real transistor are not exactly equidistant lines. In fact, they are slightly closer together for low values of base current than they are for higher values. As you can also see in *figure 2b*, the base current is not linearly related to the base voltage. This means that equal positive and negative variations of the base voltage will result in larger positive variations of base current than negative variations. Consequently, a sinusoidal variation of the base voltage will result in a collector current which alternates about its mean value, but has a slightly greater positive amplitude than negative amplitude. The output wave will therefore be a distorted version of the input. This distortion, of course, must be kept to a minimum.

We haven't worried too much about distortion before because the signal amplitude in voltage amplifiers is usually extremely small, and the transistor was not operated in the region where the nonlinearity is significant. However, power amplifiers are usually used to their full limits, and to obtain the maximum power which the device can handle, it is usually driven right across the entire set of collector characteristics – including the region where the collector is not linearly related to the base current.

Second harmonic distortion

A distorted collector current wave in which the positive and negative halves are not identical can be represented by a perfect sinusoidal wave of magnitude A_1 at the same frequency as the input, and a current of amplitude A_2 at a frequency double that of the input. The instantaneous collector current can be written as:

$$i_c = A_1 \cos \omega t + A_2 \cos 2\omega t$$

Figure 3 illustrates i_c. The black curve shows the **fundamental** component of the current – a sinusoid of amplitude A_1. The dotted curve represents the **second harmonic** component – a sinusoid of amplitude A_2. The red curve is the sum of A_1 and A_2.

The magnitude of these two collector current components can be found directly from the collector and base characteristics, by noting the maximum, quiescent and minimum currents, I_{C3}, I_{C2} and I_{C1}, that flow in response to a sinusoidal input voltage. These give us:

$$A_1 = \frac{I_{C3} - I_{C1}}{2}$$

$$A_2 = \frac{I_{C3} + I_{C1} - 2I_{C2}}{4}$$

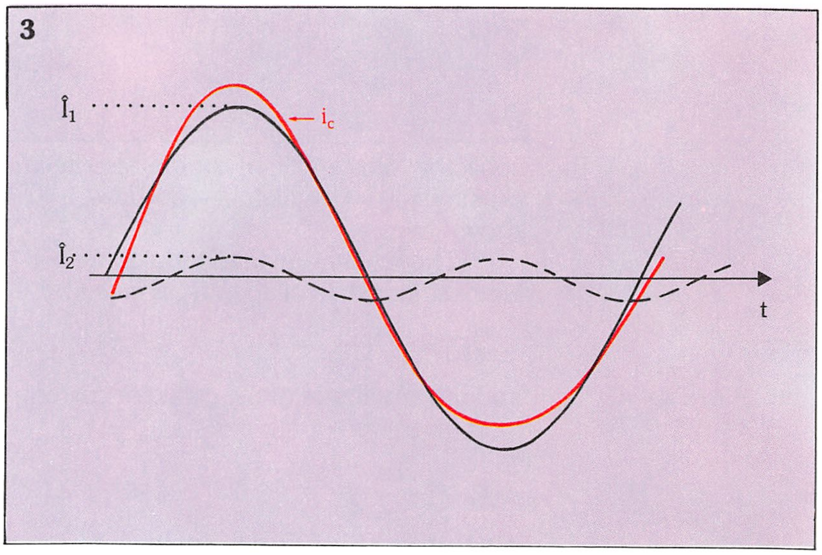
If we measure the steady current flowing in the collector circuit, with no input signal applied (i.e. the quiescent current) then we shall have a value of I_{C2}.

However, if a sinusoidal base voltage is applied this value will increase slightly to $I_{C2} + A_2$. This change in the direct current can therefore be used as a direct measure of the **second harmonic distortion**.

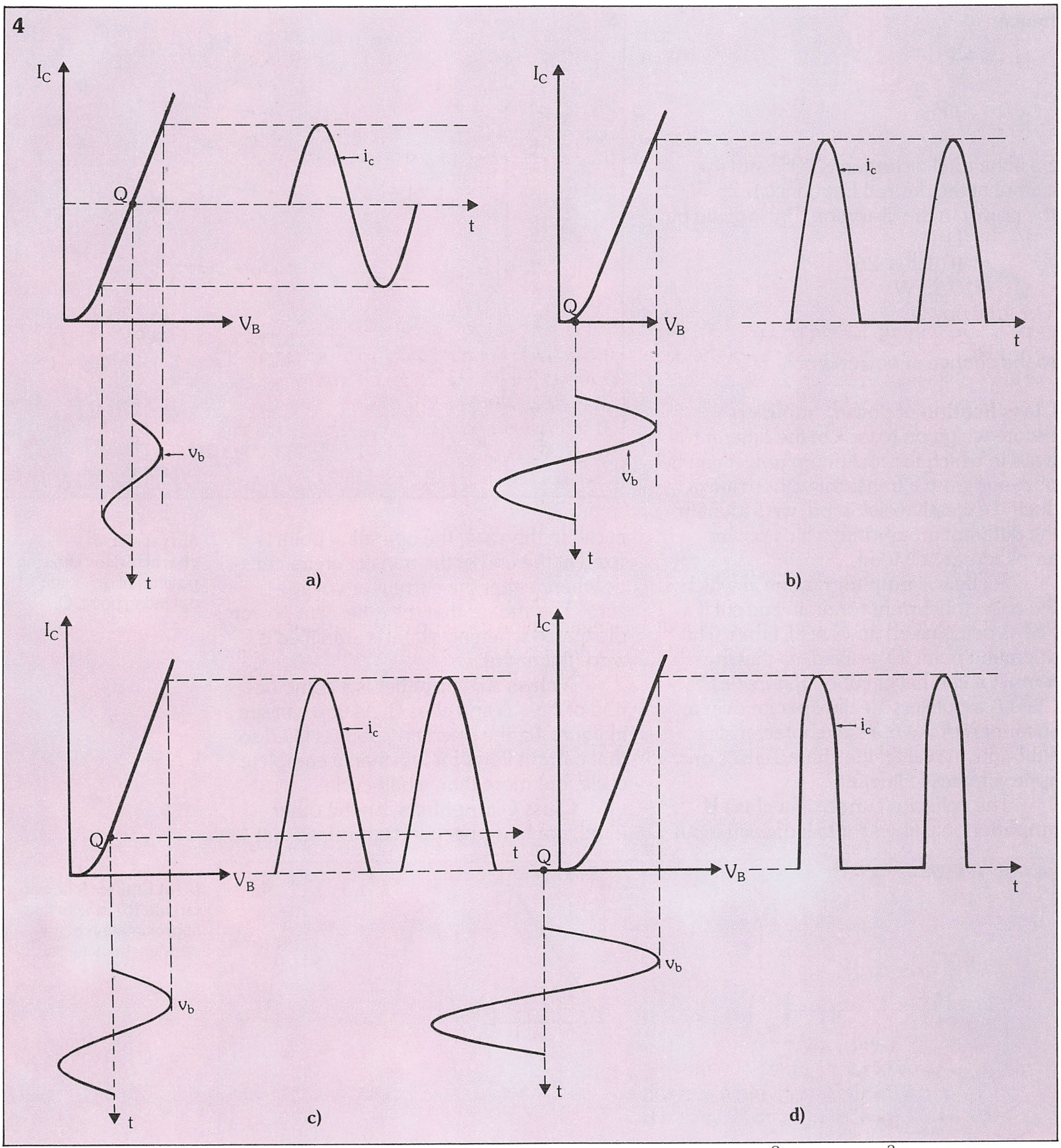
Total distortion

Power amplifiers operating in the nonlinear region develop other harmonics – of higher orders than the second. The magnitude of these harmonics is usually small when compared with the second harmonic, but the subjective effect of higher harmonics in an audio amplifier is usually more irritating.

Frequently, the magnitudes of all the harmonics are added together giving the **total distortion** in the output. If P is the total power output of the amplifier, P_1 is the desired power at the fundamental



3. Distorted collector current, i_c.



4. Input and output curves shown on transfer characteristics of: **(a)** class A; **(b)** class B; **(c)** class AB; and **(d)** class C amplifiers.

frequency, and P_D is the total distortion power at higher harmonics, we will have:

$$P = P_1 + P_D$$

The distortion, D, can be defined as the ratio of the total distortion voltage, to the total desired voltage. Now:

$$P_1 = \frac{A_1^2 R_L}{2}$$

$$P_D = \frac{A_2^2 R_L}{2} + \frac{A_3^3 R_L}{2} + \ldots.$$

where A_2 and A_3 are the amplitudes of the second, third, etc. harmonics.

If we consider A_D to be the effective voltage due to all the distortion, we have:

$$P_1 = \frac{A_D^2 R_L}{2}$$

hence:

$$D = \frac{A_D}{A_1}$$

$$P_1 = \sqrt{\frac{P_D}{P_1}}$$

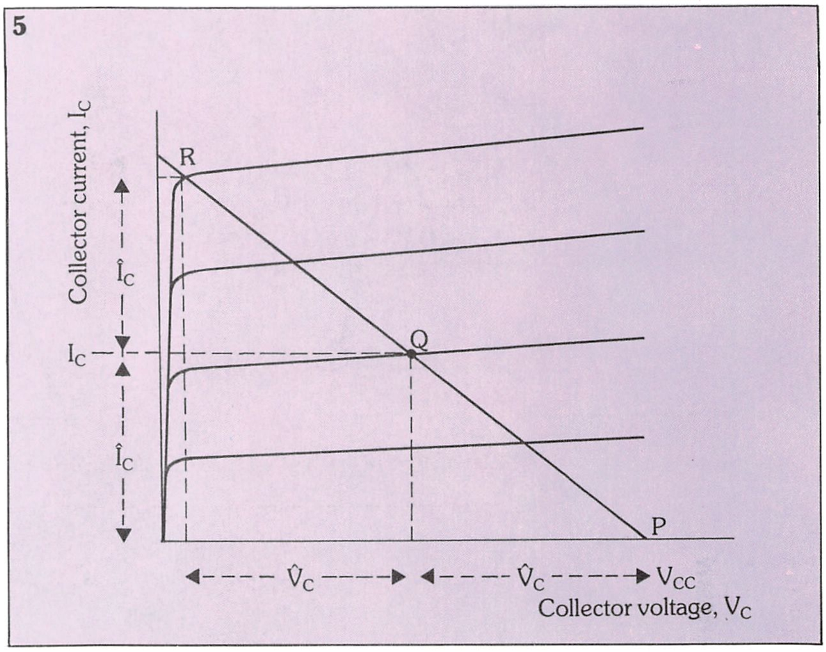
So if the total distortion is 10% and the output at the desired frequency is 25 W, the power in the distortion, P_D, is given by:

$$\begin{aligned} P_D &= D^2P_1 \\ &= (0.1)^2 \times 25 \\ &= 0.25\text{ W} \end{aligned}$$

The total power

$$P = 25 + 0.25 = 25.25\text{ W}$$

so the change in power is only 1%.



5. Transfer characteristic showing position of quiescent operating point, Q.

Classification of power amplifiers

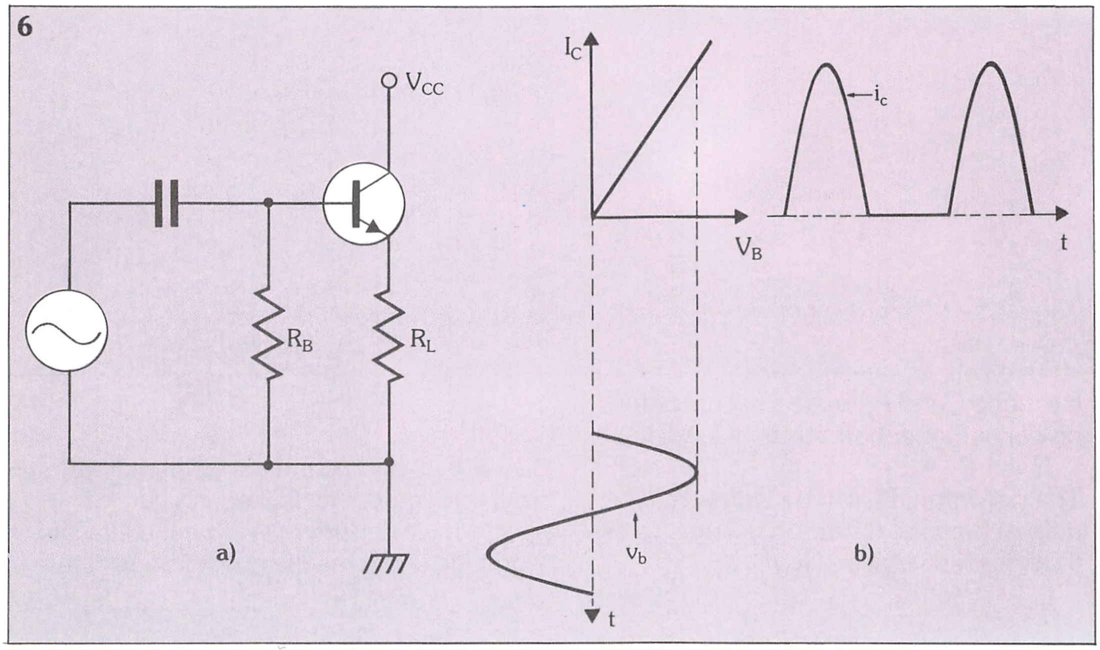
Before we go on to look at the different ways in which the maximum power can be obtained from a transistor without introducing excessive distortion, we'll identify the different groups into which power amplifiers are divided.

A **class A amplifier** is one in which the collector current (or drain current if a FET is being used) flows at all times. The operating point, Q, is fixed, so that the transistor can never cut off (*figure 4a*). Class A amplifiers ideally operate over a linear part of the transfer characteristic, although, in reality, the characteristic only approximates to linearity.

The collector current in a **class B amplifier** only flows for half of each input cycle. In this case, the operating point is fixed at the end of the transfer characteristic where either the current or voltage is zero. This means that the quiescent power dissipated when no signal is amplified is zero (*figure 4b*).

A **class AB amplifier** is a combination of class A and class B. As you can see in *figure 4c*, the operating point is fixed so that current flows for less than a complete cycle, but more than a half-cycle.

Class C amplifiers, on the other hand, are biased beyond cut-off, so that



6. (a) Emitter follower circuit; (b) straight line approximation of the transfer characteristic.

current flows for less than half the cycle of the input wave (*figure 4d*).

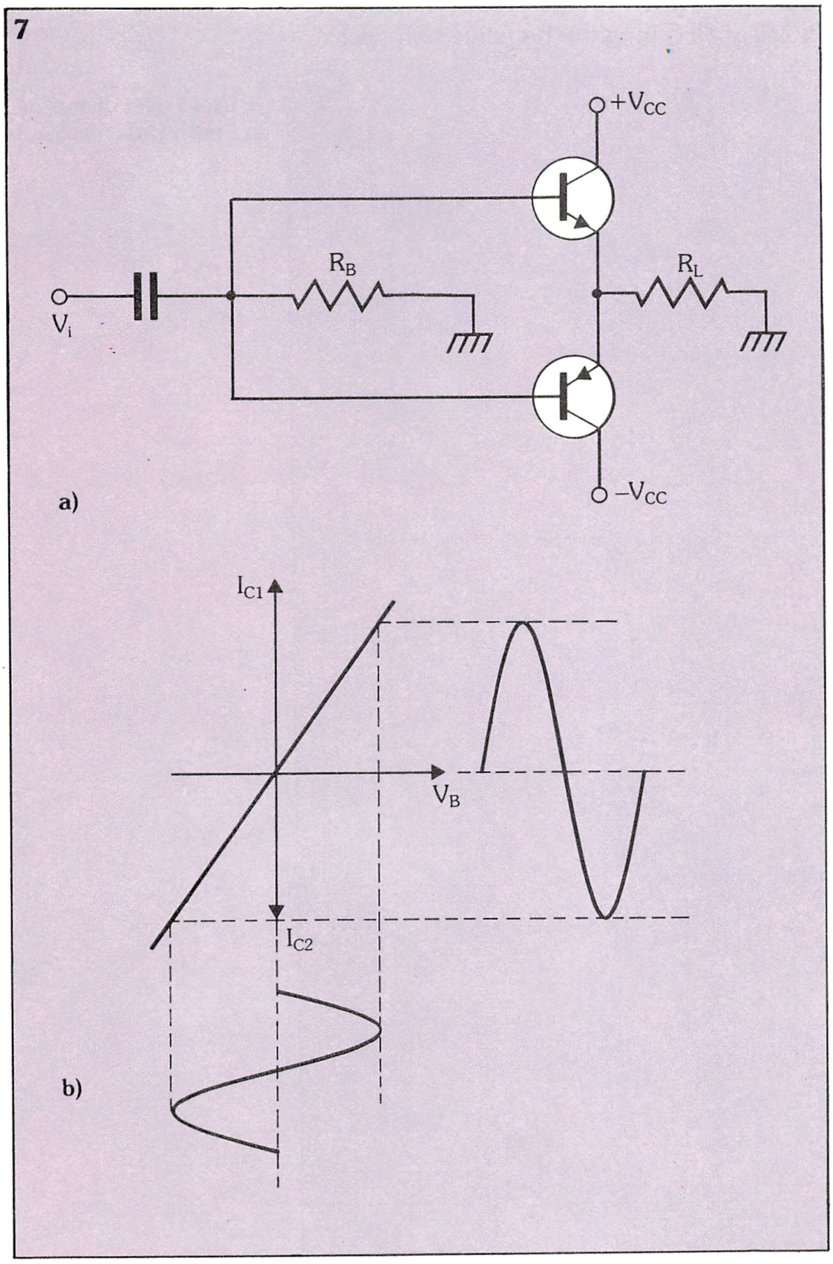
Efficiency

Maximum efficiency is desirable in any device and this must be taken into consideration when designing power amplifiers. We know that the efficiency of some devices is defined as the ratio of the total output power to the total input power. In the case of a power amplifier, the output power is the signal power delivered to the load device (a loudspeaker, for example) and the input power comprises the power supplied in the form of direct current to the collector and base, together with the signal power supplied to the base.

However, it is more convenient to use the **conversion efficiency**, otherwise known as the **collector efficiency**. This is defined as the ratio of signal power delivered to the load, to the direct power fed to the collector from the power supply. The conversion efficiency will obviously be slightly greater than the total efficiency because the signal power and the DC power supplied to the base are not included in the calculation. This doesn't mean that the conversion efficiency isn't a useful measure, as these last two factors are small compared to the collector power.

7. (a) Push pull amplifier; (b) its transfer characteristic.

7



The conversion efficiency of a class A amplifier can be found as follows. The useful power delivered to the circuit is given by:

$$P_L = \frac{A_I^2 R_L}{2}$$

$$= \tfrac{1}{2}\hat{V}_C \hat{I}_C$$

If we assume the distortion is negligible, then the DC power supplied to the load is:

$$P_O = V_{CC} I_C$$

The conversion efficiency (which is usually expressed as a percentage) is then given by:

$$\frac{\tfrac{1}{2}\hat{V}_C \hat{I}_C}{V_{CC} I_C} \times 100 = 50 \; \frac{\hat{V}_C \hat{I}_C}{V_{CC} I_C}$$

If the amplifier provides a small output signal, the excursion of the voltage, $\hat{V}_C$, and current, $\hat{I}_C$, will be small compared with the available values V_{CC} and I_C. The conversion efficiency will consequently only be two or three per cent. To obtain maximum efficiency from the transistor, we need to obtain as large a swing of voltage and current as possible. To do this, the quiescent operating point, Q, must be located on the load line midway between the cut-off condition, P, and the point, R, at which bottoming begins (*figure* 5). This increases the signal input, so that the collector current swings between the cut-off and bottoming conditions. It therefore follows that:

$$\hat{I}_C = I_C$$

Since the bottoming voltage at P is only about 0.1 or 0.2 V, we can say that:

$$\hat{V}_C = \tfrac{1}{2} V_{CC}$$

which allows us to find the conversion efficiency, η, of an ideal class A amplifier designed for maximum efficiency as being:

$$\eta = 25\%$$

As you can imagine, a conversion efficiency of 25% is not particularly desirable, as 75% of the input power is being lost as heat in the transistor, which has to be dissipated.

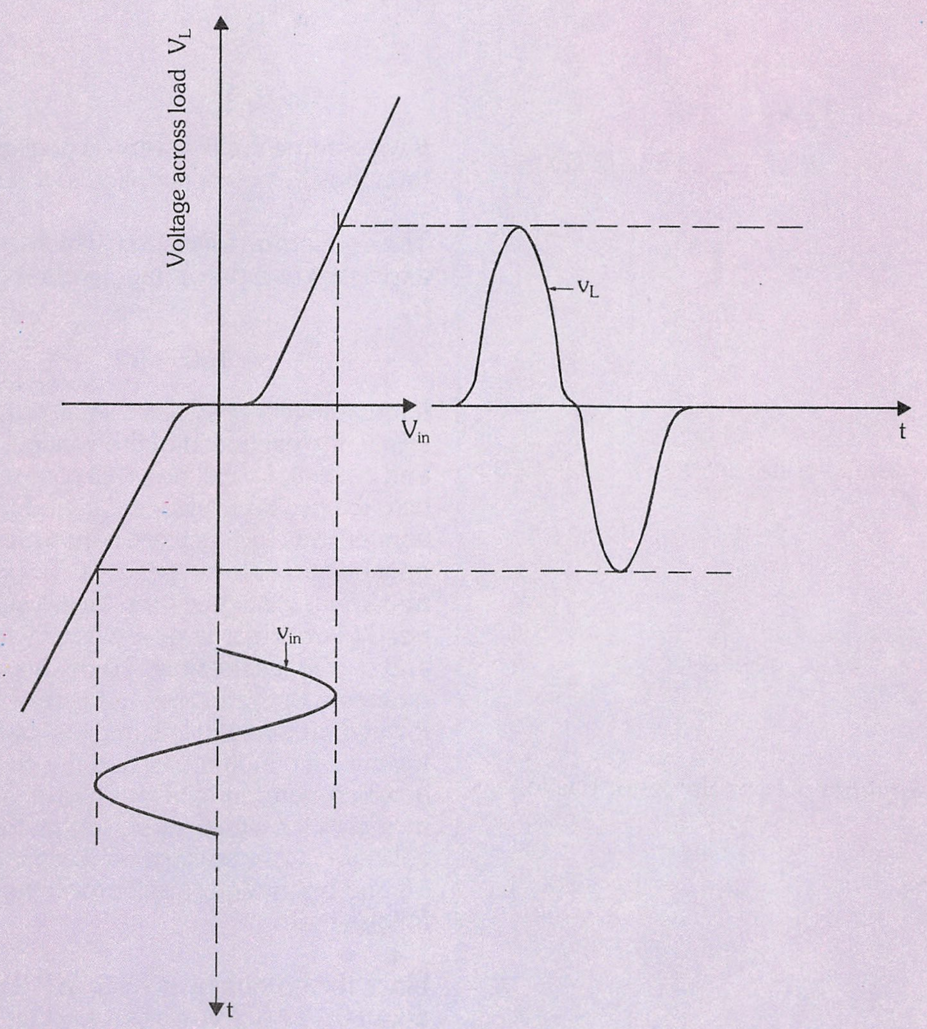
The amplifier that we have been discussing has used the transistor in the commn emitter mode, and this is usually done when the load impedance is fairly high. The common collector (emitter follower) configuration can be used when the load resistance is low – say 10 Ω or less. This design is very similar and the optimum efficiency is a gain limited to 25%.

Class B amplifier

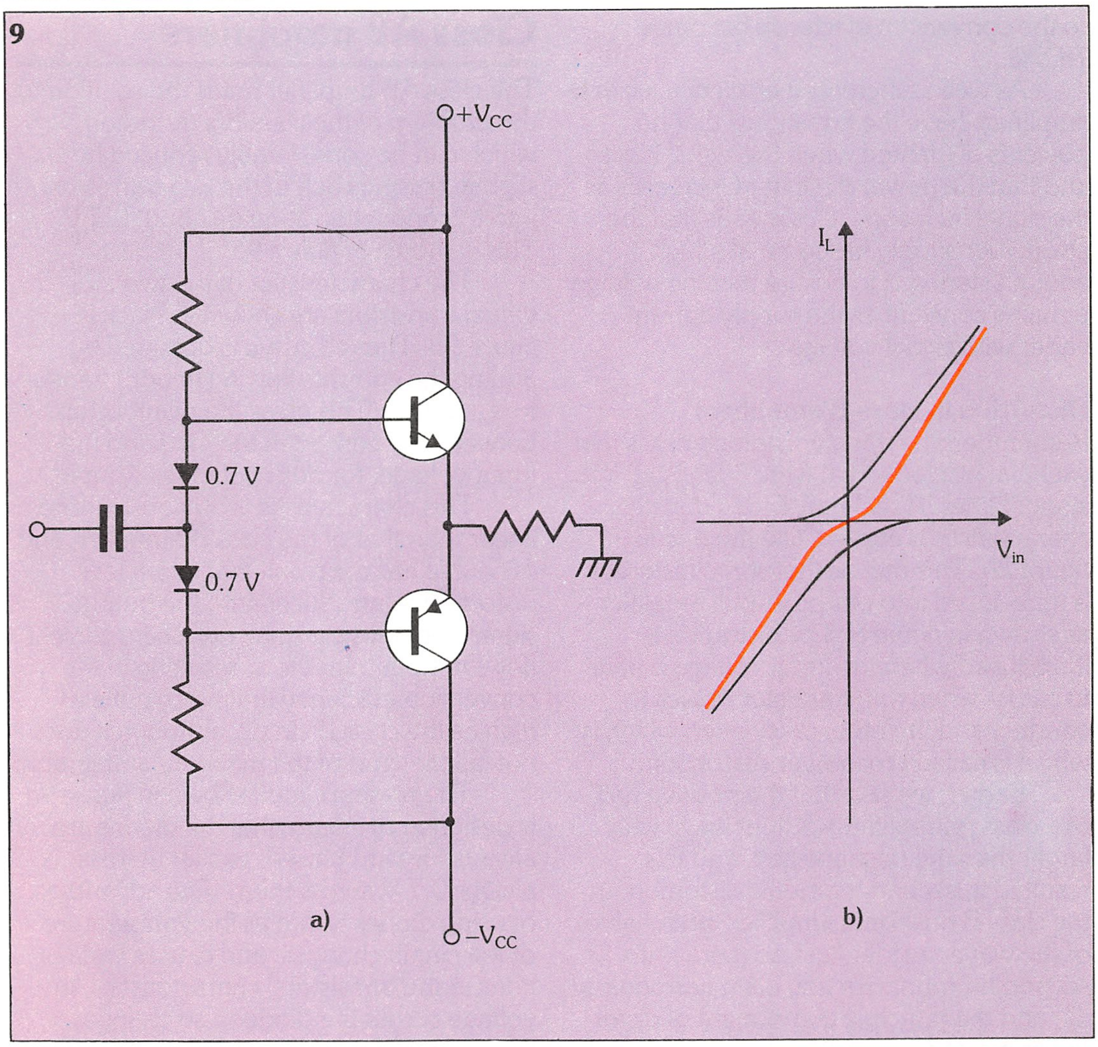
As collector current doesn't flow all the time, a class B amplifier would produce an excessive and unacceptable amount of distortion in the output circuit. However, it will be useful to take a quick look at this circuit's operation in order to find a way around the problem of distortion.

Figure 6a illustrates the more frequently used emitter follower circuit, while *figure 6b* shows an idealised straight line approximation of the transfer characteristic. As the input voltage varies sinusoidally, the output current flows in pulses only during the positive half-cycles; no current flows at all during the negative half-cycles.

8



8. Cross-over distortion in a push-pull amplifier.



9. (a) Class AB amplifier; (b) transfer characteristics of its two transistors.

However, a class B amplifier without distortion can be constructed in the following way.

A second circuit powered by a negative supply and using a p-n-p transistor conducts only during the input's negative half-cycles. The sum of the outputs from these two circuits gives an undistorted replica of the input voltage. The final circuit and its transfer characteristic are shown in *figure 7*. This circuit is known as a **push-pull amplifier**, as one transistor pushes the current up and the other pulls it down.

We can now determine the conversion efficiency for this circuit (remember, though, that this is an idealised situation). The signal power output is given by:

$$P_L = \tfrac{1}{2}\hat{V}_C \hat{I}_C$$

where:

$$\hat{V}_C = R_L \hat{I}_C$$

Power is delivered from the positive half of the supply only during the positive input half-cycles, and has an average value of:

$$\frac{V_{CC}\hat{I}_C}{\pi}$$

As a similar average power is delivered by the negative half of the supply during the negative input half-cycles, the total (average) DC power is:

$$P_O = 2\frac{V_{CC}\hat{I}_C}{\pi}$$

The conversion efficiency is therefore:

$$\eta = \frac{P_L}{P_O}$$

$$= \frac{\pi\hat{V}_C}{4V_{CC}} \times 100$$

$$= 78.5 \times \frac{\hat{V}_C}{V_{CC}} \quad \%$$

Now, if each transistor is used so that the collector voltage is driven from V_{CC} (when the input is zero) to the bottoming value (at maximum input), we will have:

$$\hat{V}_C = V_{CC}$$

so the conversion efficiency becomes 78.5%.

As well as increased efficiency, class B amplifiers have the advantage that no power is dissipated when the signal is zero, and that the power dissipated increases as the signal increases. This means that the DC power supply has to be of a high enough standard to ensure that the voltage remains constant as the supply current varies with signal voltage.

Distortion in class B amplifiers

Remember, the transfer characteristics that we have just looked at were 'idealised' into straight lines. A real transistor's transfer characteristic is curved, like the one in *figure 2b*. This means that some distortion is introduced into our push-pull amplifier – as shown in *figure 8*. The distortion is introduced when the input voltage is near to zero – when one transistor ceases to conduct and the other takes over, which is why it is called **cross-over distortion**.

Earlier, we saw that the positive half of a class A amplifier's output wave was larger than the negative half, and this resulted in second harmonic distortion. In the class B push-pull amplifier, both halves of the wave are identical, so there is no second harmonic (or any *even* harmonic at all) and the principle component of distortion is caused by the *third* harmonic of amplitude A_3. Class B amplifiers are not frequently used because of this serious distortion.

Class AB amplifiers

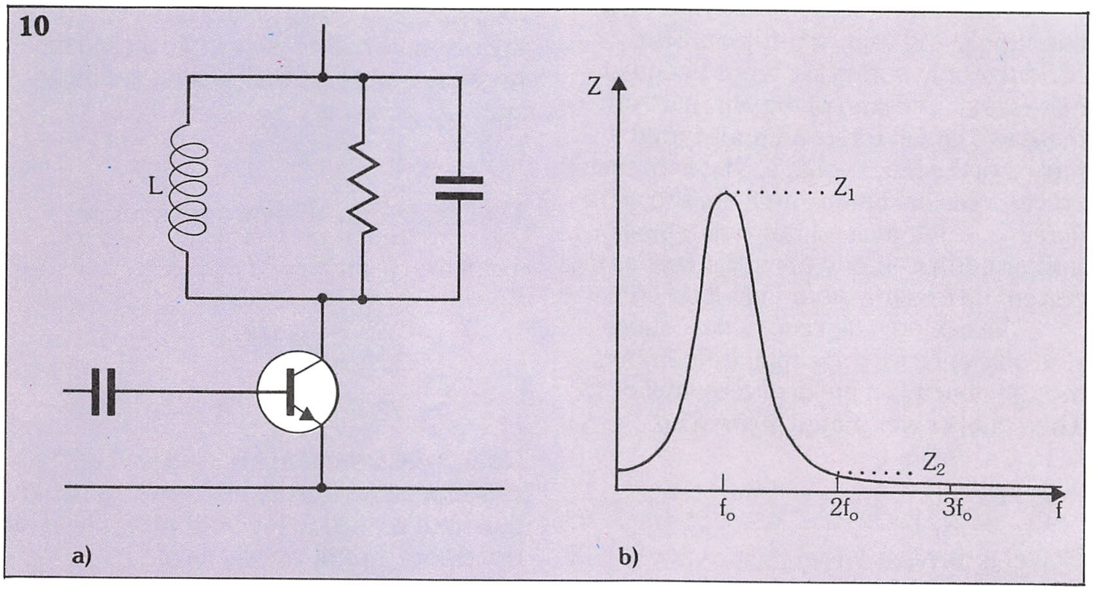
The class AB amplifier holds the solution to the problem of the class B's distortion, which can be considerably reduced by slightly biasing each of the two transistors into its conduction zone by about 0.7 V. This is shown in *figure 9a*.

The characteristics of the two individual transistors are shown in black in *figure 9b*. The red curve is obtained by adding the currents flowing in both transistors together. This gives the relationship between the current in the load and the input voltage, for both polarities of input.

This characteristic is obviously more linear than that of the class B amplifier, and of course has the result that much less distortion is introduced into the output signal. However, a small quiescent current flows in both transistors, resulting in a conversion efficiency that is not quite as high as the class B circuit, although it does not fall far short of the maximum attainable.

The biasing circuit shown in *figure 9a* is designed to ensure that the diodes are always forward biased, so that there is always 0.7 V across them. The advantage of using diodes is that as the temperature of the circuit changes, and causes variations in the transistors' characteristics, the voltage across the diodes also changes, thus correcting the bias points so that the transfer characteristic remains as linear as possible. This type of output stage is used in many op amps – like the 741.

10



10. (a) Class C amplifier; (b) plot of impedance against frequency for the tuned circuit.

Class C amplifiers

Current flows in a class C amplifier for less than half a cycle, so even if a push-pull arrangement is adopted, distortion can never be eliminated or reduced to an acceptable level. Consequently, class C amplifiers are never used to amplify speech or video signals in a linear manner. They are, however, useful as radio frequency amplifiers. In this situation, we need a circuit that provides a sinusoidal output, for a sinusoidal input with the maximum conversion efficiency. Radio frequency amplifiers are required to increase the power of a signal at one known frequency, or over a very narrow band of frequencies.

We have previously found that a distorted wave is composed of a number of frequencies consisting of the fundamental, the second harmonic and the higher harmonics. If we were to look at the waveform of the current in a class C amplifier, we would see that there will be a component at the frequency of the input, and the magnitude of this component depends on the input voltage.

If the load impedance consists of a parallel tuned circuit (*figure 10a*), then we know that at the resonant frequency, f_o, the magnitude of the circuit's impedance will be large, while at higher or lower frequencies it will get progressively smaller (*figure 10b*). If the frequency of the input signal is chosen to be at f_o, the impedance will be maximum, Z_1, and the magnitude of the voltage across the tuned circuit, V_1, will be given by:

$$V_1 = Z_1 A_1$$

where A_1 is the amplitude of the fundamental component of the current. At the frequency of the second harmonic, $2f_o$, the magnitude of the circuit's impedance will be Z_2 – very much smaller than Z_1 – and the amplitude, V_2, of the second harmonic component of the voltage will be given by:

$$V_2 = Z_2 A_2$$

Since both Z_2 and A_2 are smaller than Z_1 and A_1, we shall also find V_2 to be much smaller than V_1; higher harmonic components will, of course, be even smaller. This means that the tuned circuit's output will be an almost perfect sine wave at the frequency f_o.

You should, however, note that the output voltage is not linearly related to the input voltage since the transistor is biased well beyond cut-off, no current will flow for any input voltages smaller than the cut-off voltage.

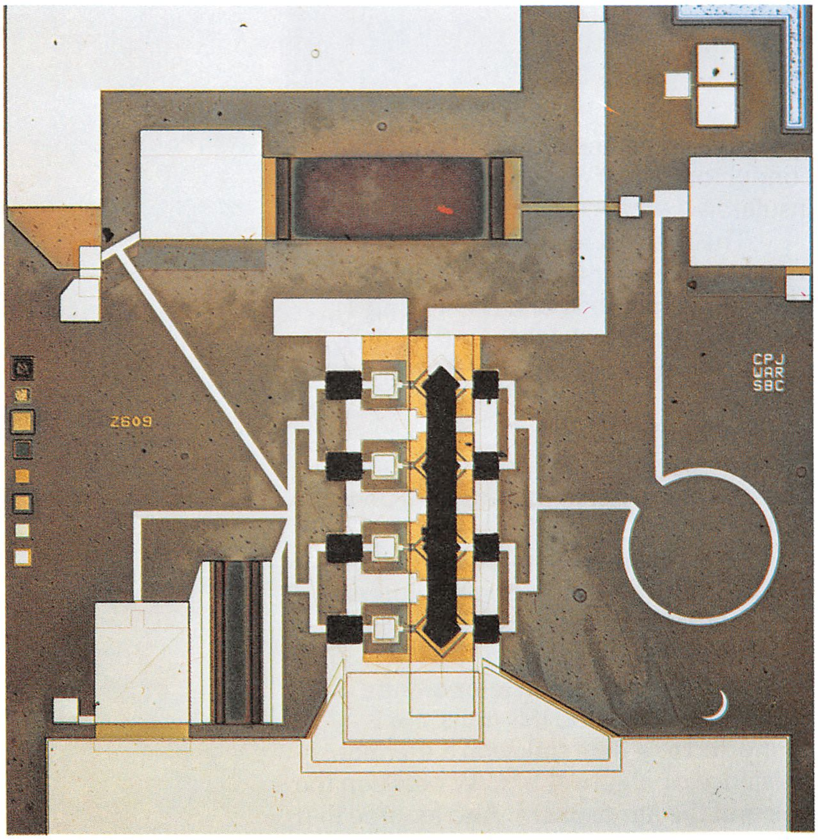
The conversion efficiency of a class C amplifier can be made very high indeed since current only flows from the supply when signal power is being drawn from the circuit.

This circuit can undergo an interesting modification if the resonant circuit is tuned to the frequency $2f_o$. This means that there will be a high load impedance at $2f_o$, and low impedance at f_o, $3f_o$, $4f_o$ etc. If the amplifier is then biased so that it only conducts for a very small portion of each cycle, the magnitude of the second harmonic of the current may be equal to that of the fundamental. In this case, the voltage across the load will principally consist of a sinusoid at frequency $2f_o$. The circuit has therefore doubled the input frequency and this is therefore known as a **frequency doubling amplifier**.

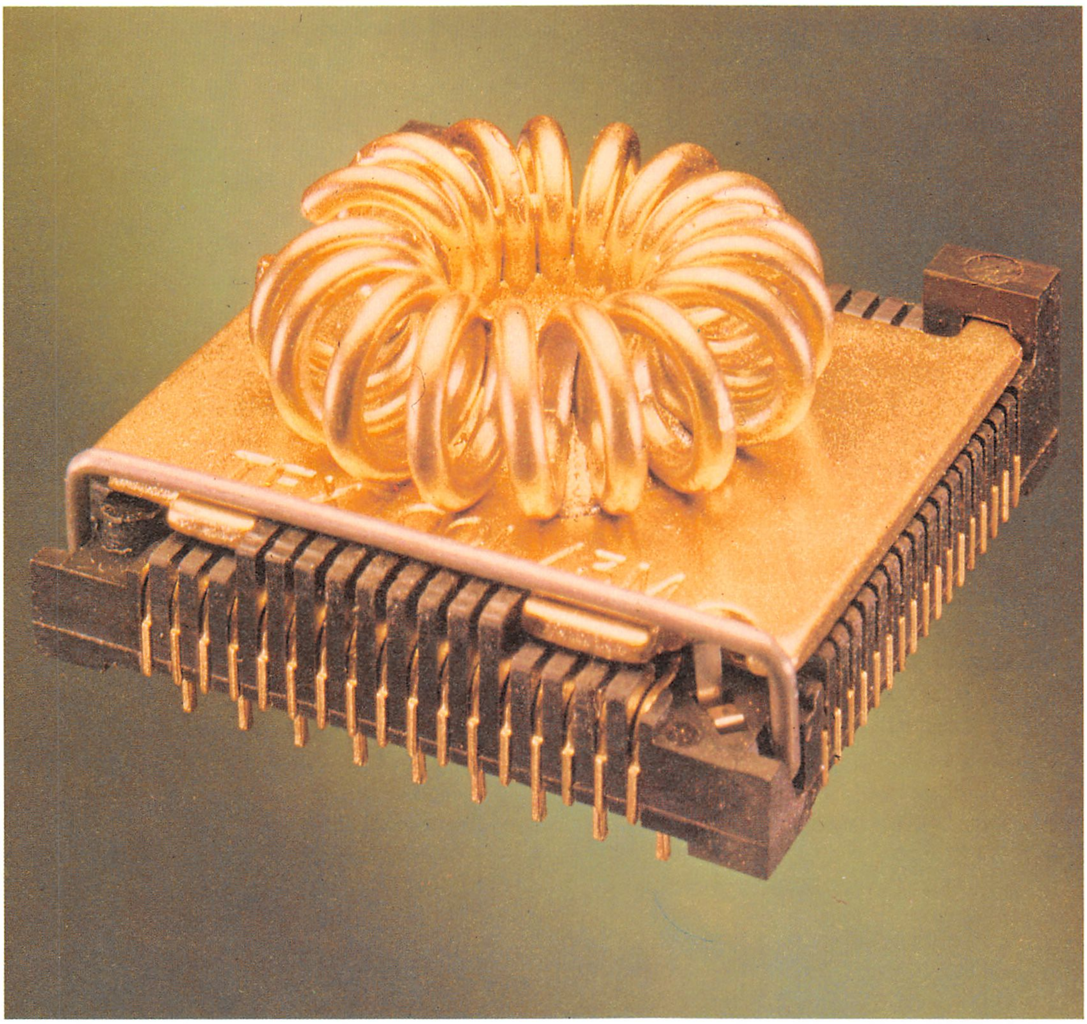
Thermal dissipation

The problem of heat dissipation, touched

Below: a 3 GHz GaAs power amplifier. A typical application would be for power output in microwave systems such as phased array radar transmitters.



The Research House/Plessey



Left: the aluminium heat sink provides a toroidal heat exchange surface – this creates high air turbulence without hindering air flow, thereby dissipating heat. (Photo: Redpoint Ltd)

on in an earlier *Solid State Electronics* article, now becomes quite significant. Large quantities of heat are generated in the transistors of a power amplifier and this needs to be conducted away. If this is not done, the transistor becomes hotter and this, in turn, causes it to dissipate more heat. If the transistor exceeds a maximum operating temperature (about 150 °C for a silicon device) it may be destroyed.

Small transistors, operating as voltage amplifiers, usually lose heat adequately by radiation to the surrounding air. Slightly larger power devices commonly have the collector mounted very close to a metal plate or finned radiator, known as a **heat sink**, which allows the heat to be dissipated more easily. Heat sinks are mounted near the collector, since the heat in a transistor is generated at the collector-base junction. Good thermal contact between the transistor and the heat sink is ensured by a film of special silicon grease which acts as a conductor of heat and as an electrical insulator.

The thermal resistance of a material is a measure of its conductivity. If the temperature at two points A and B is T_A and T_B, and the thermal resistance is θ_{AB}, then the power, P_D, which is dissipated from A to B, is given by:

$$\begin{aligned} T_A - T_B &= \Delta T_{AB} \\ &= P_D \ \theta_{AB} \end{aligned}$$

Thermal resistance is measured in units of °C/W.

As an example, lets look at the type 2N5671 transistor. The manufacturers state that the thermal resistance between collector junction and case, $\theta_{JC} = 1.25$ °C/W and the maximum temperature is 200 °C. Let's assume this transistor is connected to a heat sink with a thermal resistance of $\theta_{SA} = 4.5$ °C/W between the sink and the atmosphere. Also assume that a thermal resistance, θ_{CS}, of 0.5 °C/W exists between the transistor case and the heat sink. The maximum power which the transistor can dissipate at its collector, if the ambient temperature is 25 °C, is found as follows.

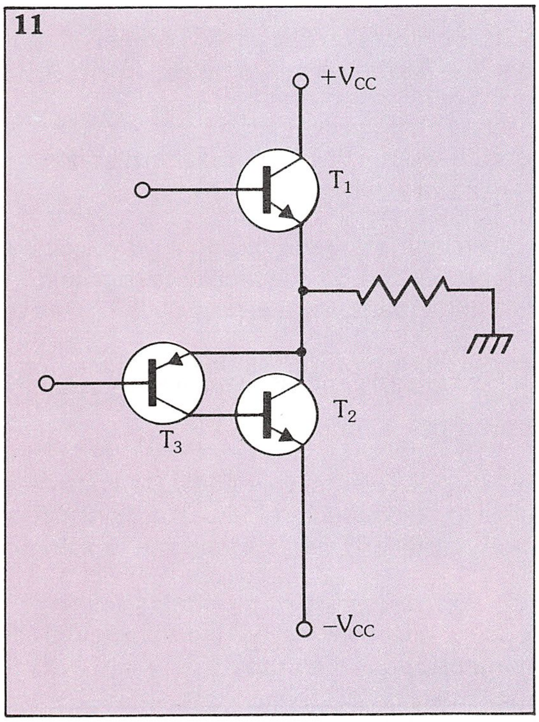
The temperature between junction and case is 1.25 P_D; between case and heat sink is 0.5 P_D; between heat sink and air is 4.5 P_D. So the temperature between junction and air is equal to:

$$1.25\,P_D + 0.5\,P_D + 4.5\,P_D = 6.25\,P_D$$

and this must not exceed 200 − 25 = 175 °C. Thus:

$$P_D = \frac{175}{6.25} = 28\text{ W}$$

So the maximum power that the transistor can dissipate under these conditions is 28 W and this occurs when the case is at a temperature of 200 − 1.25 × 28 = 165 °C. Note that the manufacturers state that this transistor is capable of dissipating up to 140 W of power if the case is kept at 25 °C above ambient temperature (in our case at 50 °C). The way that the power handling capabilities of a transistor are derated in this way is highly significant in power amplifier design and is seen to be completely dependent *on the design of the heat sink.*



11. Output circuit of the RCA type CA 2002 IC amplifier.

Integrated circuit power amplifiers

A number of power amplifiers are available as integrated circuits and we'll look at one of these, the RCA type CA 2002, in detail here. This is a relatively small amplifier capable of delivering 8 W into a load impedance of 2 Ω and may be used with loads as low as 1.6 Ω. The package contains nearly 50 transistors.

The complete circuit consists of a differential voltage amplifier stage which has two input terminals (inverting and non-inverting). This is followed by further amplifying stages, before driving the output stage consisting of a conventional class AB push-pull amplifier.

One modification in the output circuit is of interest and this is shown in *figure 11*. The lower of the two transistors in the output stage consists not of a single p-n-p transistor but of a pair of transistors: T_3, a p-n-p transistor, and T_2, an n-p-n transistor. This combination acts together like a single p-n-p transistor in a very similar way to a Darlington pair. The advantage of this arrangement is that it is difficult to obtain matched pairs of complementary transistors at relatively high power levels, and the power handling capability of p-n-p transistors is more restricted, whereas it is simple to make T_1 and T_2 as identical transistors.

The complete package is a useful power amplifier for audio signals, and has a bandwidth of up to 25 kHz. It should be used with a heat sink, the thermal resistance between junction and case being 4 °C/W and the maximum operating temperature of the collector junction must not exceed 150 °C. It has very low cross-over distortion.

The circuit incorporates a number of protection devices to ensure that the IC is not permanently damaged if the junction temperature becomes excessive or the ambient temperature is too high; such conditions merely cause a reduction in output power. It also contains circuitry to protect the device against surges in the power supply up to 40 V. The conversion efficiency at maximum power output is 58%, rising to 68% at two-thirds of the full load.

Even larger power amplifiers are available. For example, the MOS 248 can deliver 120 W continuous power into an

8 Ω load. It is constructed with an integral heat sink and large radiating fins. It uses MOSFET transistors in the output stage and operates from a supply voltage of ±55 V. The bandwidth is approximately 100 kHz and this makes it ideal for high power hi-fi systems.

Wideband operational amplifiers

Most of the op-amps we have looked at so far have been frequency compensated to have very narrow bandwidths without feedback, so that they remain stable when feedback is applied to produce a simple inverting amplifier with a gain of −1. The open loop bandwidth is frequently of the order of 10 Hz and the low frequency gain is about 10^5.

A wideband amplifier for use on video signals needs to have a bandwidth which may extend up to 6 MHz or more. To obtain this, we must accept a considerable reduction in gain. This may be obtained by external feedback resistors, but in some types of operational amplifier the feedback resistors are included with the integrated circuit, ensuring that the amplifier remains stable. Voltage gains of about 50 may be obtained with a bandwidth of up to 50 MHz.

The CA 3011 is a commercial wideband op amp and consists of the two (customary) stages of differential amplifier, with negative feedback provided within the integrated circuit between the output of the second stage and the inverting input of the first stage. This permits a typical voltage gain of 70 dB to be maintained with a bandwidth from DC up to approximately 4 MHz.

Glossary

class A amplifier — one in which collector current flows at all times. The operating point is fixed so that the transistor can never cut off and the amplifier ideally operates over a linear part of the transistor's transfer characteristic

class B amplifier — one in which the collector current only flows for half of each input cycle. Separate transistor stages are used to amplify the negative and positive halves of the input waveform, producing negative and positive outputs which are then added together

class AB amplifier — a combination of a class A and a class B amplifier, in which the output current flows for less than a complete cycle, but longer than a half-cycle

class C amplifier — amplifier biased beyond cut-off, so that current flows for less than half the cycle of the input wave

conversion efficiency — or collector efficiency. The ratio of signal power delivered to the load, to the direct power fed to the collector from the supply. Usually expressed as a percentage, represented by the greek symbol η(eta)

power amplifier — an amplifier, usually the output stage which will, for instance, deliver power – voltage *and* current – into a load, as opposed to providing a simple voltage amplification

thermal dissipation — the transistors in power amplifiers generate considerable amounts of heat, which has to be conducted away. Heat sinks allow this heat to be dissipated more easily and can prevent malfunction or destruction of the transistor

total distortion — the magnitude of the sum of all the harmonics developed in an amplifier and measured in the output